CAPITULO EXTRAÍDO DEL LIBRO:
BEING NO ONE, de
THOMA METZINGER.

Preliminary Answers

8.I The Neurophenomenological Caveman, the Little Red Arrow, and the Total Flight Simulator: From Full Immersion to Emptiness

Has a general picture of the conscious human mind emerged from the investigations carried out in the previous chapters of this book? In the end, what is it that the self-model theory of subjectivity (SMT) has to tell us about consciousness, the phenomenal self, and the first-person perspective? In this final chapter we take stock and attempt to draw the different threads of our discussion together. In the first section I offer three metaphors that will serve as introductory illustrations and increase the intuitive plausibility of the overall picture now emerging. In the next section I keep the promise I made in the first chapter: We proceed to look at some potentially new answers to those specifically philosophical questions outlined in the corresponding second part of chapter I. In the third section I draw some more general conclusions, and explore possible future routes of research. But let's start with the metaphors mentioned aboye. Each of these metaphors highlights a different set of aspects characterizing the SMT if we view it as the general outline for a theory of consciousness, the phenomenal self, and the first-person perspective. The first metaphor is of an epistemological nature, while the second and third metaphors are representationalist and technological metaphors. Each is an image of what it means to be a conscious human being. In their own way these metaphors also reflect three different and successive stages in Western history of philosophy of the mind.

The first metaphor relates to the Book VII of Plato's Republic, and the image of the cave (Plato 2000, p. II9ff.). I claim that in terms of the deeper structure underlying our phenomenal experience of the world and of ourselves in particular, we resemble neurophenomenological cavemen. The cave in which we live our conscious life is formed by our global, phenomenal model of reality. According to Plato, the cave in which we live our lives is a subterranean location, which, however, has an entrance stretched upward corresponding to the expanse of the cave and open to the light over its entire width. Our conscious model of reality is subterranean in that it is determined exclusively by the intemal properties of our central nervous system: there is a minimally sufficient neural correlate for the content of consciousness at any given point in time. If all properties of this local neural correlate are fixed, the properties of subjective experience are fixed as well. Of course, the outside world could at the same time undergo considerable changes. For instance, a

disembodied but appropriately stimulated brain in a vat could phenomenologically-enjoy exactly the same kind of conscious experience as you do right now while reading this book. In principle, it would even suffice to properly activate just a subset of this brain, the minimally sufficient neural correlate of your present state, to make a "phenomenological snapshot" of exactly the same kind of conscious experience emerge. Of course, we would never call such a physically restricted phenomenal subject a person, or even a subject in any philosophically interesting sense at all. For example, such a sub- personal clone of your own current conscious model of the world right now would bizarrely misrepresent its own position in the world; it would have an extreme number of false beliefs about itself. There would be experience, but not knowledge. Still, it is true lo say that phenomenal experience as such unfolds in an internal space, in a space quite dis, tinct from the world described by ordinary physics. It evolves within an individual model of reality, in an individual organism's brain, and its experiential properties are determined exclusively by properties within this brain. Although this simple fact may well be cognitively available to many of us, we are neurophenomenological cavemen in that none of us are able to consciously experience its truth. Effortlessly, we enjoy an "out-of-the-brain experience.` Only if confronted with the data and discoveries of modem neuropsychology, or if pressed to come up with a convincing argument showing that currently we are not just a shadow on the wall of the phenomenal cave generated by some sort of isolated, minimally sufficient correlate stimulated by an evil scientist, only then do we sometimes begin lo develop a stronger intuitive sense of what it means that our phenomenal model of reality is an internal model of reality that could at any time, in principie, tum out to be quite far removed from a much more high-dimensional physical reality than we have ever thought of. Plato, however, tells us there is an entrance to the cave, which at the same time may be a potential exit. But who could it be? Who could ever pass through this exit?

In Plato's beautiful parable the captives in the cave are chained down by their thighs and necks. They have been in this position since birth, and they can only look straight ahead, because even their head has been in a fixed position from the beginning of their existence onward. They are prevented by their fetters from tuming their heads. As Socrates points out, they have never seen anything of themselves and each other except the shadows cast by the fire burning behind them to the opposite wall of the cave, and which they take for real objects. The same is true of the objects carried along aboye the low wall behind their heads. What is the cave? The cave, according to the SMT, is simply the physical organism as a whole, including, in particular, its brain. What are the shadows on the wall? A shadow is a low-dimensional projection of a higher-dimensional object. Phenomenal shadows are low-dimensional projections of internal or external objects in the conscious state space opened within the central nervous system of a biological organism. According lo the SMT, the shadows on the wall are phenomenal mental models. The book you are

holding in your hands, as consciously experienced by you at this moment, is a dynamic, low-dimensional shadow of the actual physical object in your hand, a dancing shadow in your central nervous system. As all neural network modelers know, real-life connectionist systems typically achieve a major reduction in the dimensionality of their input vectors at the very first processing stage, when transforming the activation pattern on their sensory surface into the first hidden layer. But what is the fire, causing the projection of flickering shadows of consciousness, ever changing, dancing away as activation patterns on the surface of your neural cave? The fire is neural dynamics. The fire is the incessant, self-regulating flow of neural information processing, constantly perturbed and modulated by sensory and cognitive input. The wall is not a two-dimensional surface. It is a space, namely, the high-dimensional phenomenal state space of human technicolor phenomenology (see McGinn I989b, p. 349; Metzinger 2000b, p. If.). Please note that, in a conscious human being, the wall and the fire are not separate entities: they are two aspects of one and the same process. But what exactly does it mean when Plato tells us that we have never seen anything of ourselves but our own shadow on the opposite wall? It means that, as perceiving, attending, thinking, and even as acting subjects, we are only given to ourselves through what I have called the PSM-the phenomenal self-model. Could we free ourselves from our attachment to this inner image of ourselves, the dancing shadow in our conscious state space? Could we stop to confuse ourselves with this shadow, and leave Plato's cave altogether?

I. The notion of an "out-of-the-brain experience" was first comed by Revonsuo 2000a, p. 65. The functionál principie of internality and the ontological principie of local supervenience are not reflected on the level of conscious experience itself, because-on the level of representational content-it is systematically "externalized": the brain is constantly creating the experience that I am directly present in a world outside my brain, although, as Revonsuo points out, the experience itself is brought about by neural systems buried deep inside the brain. See also sections 7.2.3 and 7.2.5.

Here is where we have to depart from the classical metaphor. I claim that there is no one in the cave. There is no one who could leave. There is no one who-in Socrates' words-could "stand up suddenly and turn his head around and walk and . . . lift up his eyes to the light" (5I5c; p. I23), who could retum to the cave, after having seen the light of the sun, the "dazzle and the glitter" (5I5c; p. I23) of true reality, and there is no one who could later provoke the laughter of the ignorant perpetual prisoners, about whom Socrates asks the following question: "And if it were possible to lay hands on and to kill the man who tried to release them and lead them up, would they not kill him?" (5I7a, p. I29).

It is important to note that a shadow, although dependent on, controlled by, and in a certain, very weak sense representing the object that casts it, is never a distinct entity. Shadows as such don't exist. What exists are shaded surfaces. However, it is, of course, possible to confuse object and shaded surface, thereby treating the latter as a distinct entity. I claim that the conscious self is not a thing, but a shaded surface. It is not an individual object, but a process: the ongoing process of shading. The beauty of the shadow metaphor for self-consciousness consists partly in the fact that it is not only a classical but also a global metaphor-one to be found at the origin of many of mankind's great philosophical traditions. To name a prominent non-Western example, Samkara (who lived I200 years later than Plato, from 788 A.D. to 820 A.D.), in his Vivekacúdámani, or Crest-Jewel o Wisdom (Samkara I966, p. 70), argued that just as we don't confuse ourselves with the shadow cast by our own body, or with a reflection of it, or with the body as it appears in a dream or in imagination, we shouldt not identify with what appears to be our bodily self right now. Samkara said: Just as you have no seif-identification with your shadow-body, reflection-body, dream-body, or imagination-body, so should you not have with the living body. The SMT offers a deeper understanding of why, in standard situations, the system as a whole inevitably does identify itself with its own neurodynamical shadow, with its inner computational reflection of itself, with its continuous online dream about, and internal emulation of, itself. It is the transparency of the human seif-model which causes this effect.

We must imagine Platós cave differently if we are to understand the neurophenome-nological caveman's trae situation. There are low-dimensional phenomenal shadows of external perceptual objects dancing on the neural user surface of the caveman's brain. So much is true. There certainly is a phenomenal self-shadow as well. But what is this shadow the low-dimensional projection of? I claim that it is a shadow not of a captive person, but of the cave as a whole. It is the physical organism as a whole, including ail of its brain its cognitive activity, and its social relationships, that is projecting inward, from all directions at the same time, as it were. There is no true subject and no homunculus in the cave that could confuse itself with anything. It is the cave as a whole, which episodically, during phases of waking and dreaming, projects a shadow of itself onto one of its many internal walls. The cave shadow is there. The cave itself is empty.

Samkara was right: A transparent phenomenal self-model is not a self. But Socrates was right too. He depicted the prisoners as fully anchored in the cave, chained down since birth. Exactly the same holds true for our phenomenal self-model: It is firmly anchored in the autonomous bodily dynamics of elementary bioregulation, through a procesa I call "self-presentationt." The human self-model transforms our lived reality into a centered reality, because it is the only phenomenal shadow firmly anchored in a continuous source of internally generated input. Socrates clearly saw that a persistent functional link was there. I return to the issue of whether this link could ever be broken when discussing the third

metaphor for the SMT at the end of this section. According to the SMT, it is true that the dancing shadow on the internal wall of our brain possesses a persistent functional link to this very brain, for example, as realized by the upper brainstem and the hypothalamus. It is not true that there is an intemal person forming the object of this shadow, a person conceived of as a distinct entity tied down by such a functional link. Personhood is a global property of the system as a whole that only emerges at a much later stage, through social interactions. The self-shadow-a necessary precondition for al] social interaction-is simply the shadow cast by the cave as a whole onto itself. Plato was also right about the extremely reduced dimensionality of our phenomenal model of reality. From all we know today, the flow of conscious experience is an idiosyncratic trajectory through phenomenal state space, a highIy selective projection shaped by the contingencies of biological evolution on this planet-something much more resembling a reality tunnel through an inconceivably high-dimensional reality. A third aspect, in which both Plato and Samkara were certainly right, is the normative ideal of expanding self-knowledge. The neurophenomenological caveman's situation is deplorable. It must be changed. However, it cannot be changed by freeing ourselves and leaving the cave altogether, searching for the true lght of the sun. We have never been in the cave. The cave is empty.

The second metaphor I want to offer here is a representationalist metaphor. Representationalist theories of mind have a long history, spanning many centuries of Western philosophy. Recently, representationalist theories of conscious experience have again become popular and the mundane concept of a "map" has at the same time become a ubiquitous tool in neuroscience and the cognitive sciences. The idea is simple and straightforward: Phenomenal experience is like a dynamic, multidimensional map of the world. Interestingly, like only very few of the external maps used by human beings, it also has a little red arrow. I claim that the phenomenal self is the little red arrow in your conscious map of reality.

When looking at a city map on the wall of a subway station you will frequently discover a little red arrow, maybe even a sentence next to it saying, YOU ARE HERE! It is interesting to note that this linguistic explanatory note is not strictly necessary. For most users, a map exhibiting only the little red arrow will serve its functional purpose just as well. Your phenomenal map of the world is an internal map of the world. In order to be useful, it must have more than phenomenal content alone-very roughly, it must possess a certain isomorphy to your current environment. This is the problem of intentionality: there must be some kind of link between the map and the city, between mind and world. In order to achieve a certain degree of covariance with external reality, it must also be a dynamic map, capable of constant, flexible, and swift updating. However, a conscious model of reality has only one single user. This is not true of the map in the subway station. The subway map has many users. It does not change with the city around it. It is an external

object. Compared with the enormous wealth of your conscious model of reality, it is less than a shadow. Not only is it a low-dimensional projection, it does not possess a genuine first-person perspective; all it has is a little red arrow. The little red arrow is the self-model of the city map user. It specifies the position and thereby, indirectly, the interests of potential users such as an external representation of reality within this representation. The little red arrow and the indexical sentence YOU ARE HERE! deprives it of its universal character and turns it finto an instrument of orientation, which can only be used successfully at a single location in the world.

The multimodal maps generated by human brains, however, are general models df reality that flexibly adapt to the situation of the organism and are'updated in real tinte. Since they are also internal models of the world, the user, whose purposes they have Lo serve, is in fact identical across all possible situations. As opposed to firmly installed maps in subway stations it is not the problem domain, which is fixed, and the class of users, which is variable, but the system as a whole, which remains identical across all representational situations while the class of problems is so general as to be almost infinite. Humean beings are general problem solvers and autonomous agents at the same time, developing a phenomenal geography of the world. Mental self-models are the little red arrows that help a phenomenal geographer to navigate her own complex mental map of reality by once again depicting a subset of her own properties for herself. As long as they are functionally active, they transform the models of reality in which they are embedded by the system as a whole into user-centered representations. Consciousness is typically tied to an individual perspective, and it is not only by reason of their physical internality but owing to their structural and representational fixation to a single user that centered models of reality ', are transformed into meaningful instruments for only a single system in the world. Insofar as their functional profile is additionally characterized by extraorganismic relations, self-models cannot even lose the property of physical internality. Not only are they anchored in a fine-grained internal context, some of their higher layers are also driven by their social environment. We therefore return to the issue of the portability of self-models below.

For now it is only important to point out that the uniqueness of every single phenome-nal nal subject is anchored in the uniqueness of the functional properties constituting the selfmodel underlying it. This self-model is the little red arrow that a human brain uses to orient itself within the internal simulation of reality it has generated. The most important difference between the little red arrow on the subway map and the little red arrow in our neurophenomenological troglodyte's brain is that the external red arrow is opaque. It is always: clear that it is only a representation-a placeholder for something else. The little red arrow on the subway map is clearly recognizable as a variable, because different passengers can use this map by identifying with this little red arrow-they are episodically, as it were, "incarnated" in the reality model constructed by the map. This representational incarnation

in external media of representation is something that could never work without a con-, scious self-model. The conscious self-model in the caveman's brain itself, however, is in large portions transparent: it is not experienced as a representational structure, not as a placeholder and not as a variable. It is a phenomenal self characterized not only by full-blown prereflexive embodiment but by the comprehensive, all-encompassing subjective experience of being situated.

Could there be an external user, someone who became deeply entangled within our current conscious model of reality by mistakenly identifying herself with the little red arrow, the PSM? Are we like moviegoers who have identified so strongly with their hero on the screen as to have completely forgotten who and where they actually are? No. The cave is empty. What the cave in internally generating a multidimensional neural image of itself as a whole allows to emerge, however, is a fascinating phenomenal property: the property of "full immersion." This property plays the central role in our third and last metaphor. As it turns out, in reflecting functionalist intuitions in the philosophy of mind, this metaphor is closest to the present time: It is a technological metaphor, and as all readers educated about current virtual reality technology will have noted, the concept of "full immersion" in its origin is a technological concept as well. As the history of philosophy has shown, technological metaphors are dangerous if their limitations are not clearly seen. Let us keep this issue in mind as we begin with a slightly old-fashioned image.

I claim that phenomenal first-person experience works like a total flight simulator. A flight simulator is a device for student pilots. It can also serve for training in behavioral reactions to unforeseen and critical situations without the risk of a real-world crash. Flight simulators were already in use at the beginning of the last century, and since then they have been continually improved. In yesterday's standard model, candidates sit in a cabin that rests on a movable platform on large extensible legs (figure 8.I). The legs are controlled by a computer that can mimic all motions of an airplane.

One of the most important practical tasks in successfully programming vehicle simulators lies in understanding the "tolerable dynamical limits in visual-vestibular miscorrelation" (Ellis I995, p. 25), because in order to create a coherent virtual environment two very different sources of sensory information have to be integrated: the proprioceptive sense of balance and the external sense of vision. The phenomenal self-model (as driven by the simulator) has to cohere to the phenomenal world-model (as driven by the simulator). In the cabin we find a cockpit of realistic design, containing all the displays and control instruments one finds in a real airplane. The student pilot views a computercontrolled video screen, supplying him with a visual simulation of the view from the cockpit. In more advanced models this screen will have been replaced by a data helmet, containing two slightly displaced monitors creating a view into three-dimensional surround graphics. It is characterized by an "infinity optics." A special programming technique serves to keep the

virtual focus of the image always at more than I0 yards distant. If the candidate looks "out of the window," he is able to focus his eyes on distant objects, although the real computer-generated image is only a few inches away from his face. This visual simulation of external reality is constantly updated at great speed depending on the actions the pilot takes. Today, it is also possible to specifically stimulate the proprioceptive and kinesthetic senses, for instance, by employing a seat shaker that helps to simulate a whole range of bodily sensations, as they are typically generated by a sudden break in airflow during critical velocities or vibrations of the afterburner. In this way, the student pilot learns how to use onboard instruments, gets to know the reactions of an aircraft to his own actions, carrying out the most important basic operations any good pilot needs to master, and without taking any major physical risks.

Human brains function in a similar way. From internally represented information and utilizing continuous input supplied by the sensory organs they construct an internal model of external reality. This global model is a real-time model; it is being updated at such a great speed and with such reliability that in general we are not able to experience it as a model anymore. Phenomenal reality, for us, is not a simulational space constructed by our brains, but in a very direct and experientially untranscendable manner it is simply the world, in which we live our lives. A fight simulator, however, is easily recognized as a flight simulator. Although as student pilots we work in it in a very concentrated fashion, we never believe we are really flying. The reason for this opacity of the artificial simulation surrounding us is simply that our brain continuously supplies us with a much better reference model of the world than the computer that controls the flight simulator. The images generated by our visual cortex are orders of magnitude faster and certainly more reliable, they are characterized by a much higher resolution and a greater wealth of detail than the images appearing on the monitor of a training simulator. This is why we can always recognize the images on the monitor as images at any point in time, simply because we possess a higher representational standard with which we can compare them. If the simulator starts to shake and rattle as the result of flying through an air pocket or the consequences of an inadvertent maneuver, then these shaking and rattling motions will never truly deceive the student pilot. This is so because the phenomenal models of our own bodily motions generated from proprioceptive and kinesthetic perceptions are much richer in detail and more convincing than the simulations of airplane movements generated by the computer ever could be. However, it must be noted, this situation will doubtlessly change soon (for an excellent overview, see Barfield and Furness I995). The subjective experience of presence and being there is determined by functional factors like the number and fidelity of sensory input and output channels, the ability to modify the virtual environment, and, importantly, the level of social interactivity in tercos of actually being recognized as an existing person by others in the virtual world (Heeter I992).

From an engineering point of view, the problems involved in creating virtual environments are problems of advanced interface design. A virtual interface is defined as a system of transducers, signal processors, hardware, and software that creates an interactive medium conveying information to the senses in the form of 3D virtual images, tactile and kinesthetic feedback, spatialized sound, and so on, while monitoring the psychomotor and physiological behavior of the user and employing it to manipulate the virtual environment (Barfield and Furness I995, p. 4). Virtual environments are the latest development in neurophenomenological cave art. And, obviously, this is one fruitful way of looking at consciousness. Phenomenal experience, insofar as it is transparent, is an invisible interface, an internal medium that allows an organism lo interact with itself. It is a control device that functions by creating an internal user surface. Moreover, if one looks at hqw theorists in virtual reality and advanced interface design today actually define the attributes of what, for them, would be an ideal medium, one is immediately reminded of the catalogue of constraints for phenomenal representations offered in chapter 3 (table 8.I).

Table 8.I

Attributes of an ideal medium: Conscious experience as an invisible interface

- *Matches the sensory capabilities of the human*
- *Easy to leam*
- *High bandwidth bridge to the brain*
- *Dynamically adapts to the needs of the task*
- *Can be tailored to individual approaches*
- *Natural semantic language*
- *Organization of spatial/state/temporal factors*
- *Macroscopic vs. microscopic view*
- *High bandwidth input*
- *Information clustering*
- *Information filtering*
- *Unambiguous*
- *Does not consume reserve capacity*
- *Easy prediction*
- *Reliable*
- *Operates when busy*
- *High semantic content (sample presentahon*
- *Localization of objects:*
    - *movement*
    - *state*

immediacy

- *Sense of presence*

From Furness and Barfield I995, reprinted with permission from Oxford University Press.

The virtual reality metaphor for phenomenal experience possesses great heuristic fer tility, but we must not lose sight of its inherent limitations. The conscious brain differ& from a flight simulator in a number of relevant aspects. First, it possesses many', modalities and presentational subformats: Just think of conscious vision, auditorY: phenomenology, olfactory and gustatory qualities, tactile sensations, and the incredibW subtlety and richness given through bodily interoceptors. In particular, it is able to.,bI the information originating in all these different modalities into a nonfragmented, UM model of reality-and it is precisely this task that even in a flight simulator is left to brain of the student pilot. Flight simulators drive phenomenal models of reality, but do not yet create them. Second, as opposed lo flight simulators and present-day v reality systems, the human brain is not confined lo a specific domain. The conscious is open lo a vast number of representational configurations and simulational tasks. As aboye, conscious brains approximate the classic notion of a general problem solver Newell and Simon I96I). A thid characteristic, however, distinguishing brains and simulators is much more important in our present context: human brains simulate the pilot as well.

Of course, there is no homunculus in the system. The cave is empty. The little red arrow is just a special representational device. What does exist for conscious systems of a certain complexity, however, is a certain need-the necessity for the system as a whole lo explain its own inner and outer actions lo itself. It has lo possess a representational and functional tool that helps lo predict its own future behavior, lo continuously monitor critical system properties with the help of an ongoing internal simulation, and which can depict the history of its own actions as its own history. Generally speaking, the system needs a computational tool that helps it in owning its own hardware. This tool is what I have described as the self-model of the organism. The brain differs from the flight simulator in not being used by a student pilot, who episodically "enters" it. It operates like a "total flight simulator": A total flight simulator is a self-modeling airplane that has always flown without a pilot and has generated a complex internal image of itself within its own internal flight simulator. The image is transparent. The information that it is an internally generated image is not yet available lo the system as a whole. Because, operating under the condition of a naive-realistic self-misunderstanding, it interprets the content of this image as a non-physical object; "the pilot" is born in its flight simulator. Like the neurophenomenological caveman, "the pilot" is bom into a virtual reality right from the beginning-without a chance lo ever discover this fact. Like a seriously deluded tourist who actually believes he is the

little red arrow, the caveman is like an airplane that functionally owns its hardware, but has only just begun lo appropriate the simulator. Neurophenomenologically, he is a shadow boxer who has become hypnotized by his own internal shadow. Employing some more recent terminology, the pilot rather is like a biologically grounded "softbot," a humanoid "avatar" used by the airplane as its own internal interface lo control its own hardware, as a whole, and more flexibly.

It is surprising to see how theorists researching virtual environments today not only employ phenomenological notions like "presence" or "situatedness" but have already comed a terminological notion for what, under the SMT, would be the spatial partition of the PSM modeling motor properties of the organism: the "virtual body" (VB; Barfield, Zeltzer, Sheridan, and Slater I995, p. 505). A VB is a part of an extended virtual environment, a dynamic and high-dimensional tool that can be used as a little red arrow. It can be used lo control a robot at a distance, employing the virtual body as an interface.

However, the authors also point out how the issue of "identification" is crucial lo the context of teleoperator systems controlling distant robots, and how users of a virtual environment may actually reject their VB, just as some neuropsychological patients do (ibid., P. 506). Most illustrative, however, is the notion of a "slave robot": To achieve telepresence, an operator has lo rely on a high correlation between his own movements as sensed "directly" and the actions of the slave robot; and he ideally has lo achieve an identification between his own body and that of the slave robot.

A VB, like a PSM, is an advanced interface to functionally appropriate and control a body. In the VB case, the body may be thousands of miles away, and the interface used will (hopefully) only be episodically transparent. In the PSM case, Mother Nature solved all major interface problems millions of years ago, including a VB and extensive internal user-modeling. The target system and simulating system are identical, and conscious subjectivity is the case in which a single organism has learned to enslave itself. Interestingly, this does not turn the system as a whole into a slave robot, but finto an increasingly autonomous agent. Autonomy is conscious self-control. However, in the early stages a price has to be paid. The representational misunderstanding then generates a phenomenal! self-misunderstanding on the level of phenomenal experience, as explained in sectionsr 3.2.7 and 6.2.6. It is phenomenal transparency, a very special kind of darkness, which generales this fundamental deficit in subjective knowledge concerning the constitutive conditions and the deep structure of our own phenomenal self-consciousness, which later leads to misguided philosophical theories like the Platonic metaphor of the helmsman or ur the homunculus in the cave, which leads to the birth of the Cartesian ego and eventually to the Kantian notion of a transcendental subject, to the many false theories of "the pilot," whose existence preceded that of the body and who only episodically "entered" into it. Shadows are a form of darkness. Growth of knowledge in cognitive neuroscience today makes all

these classical models look untenable. On the contrary, the brain, the dynamical, self-organizing system as a whole, activates the pilot if and only if it needs the pilot as a representational instrument in order to integrate, monitor, predict, and remember its own activities. As long as the pilot is needed to navigate the world, the puppet shadow dances on the wall of the neurophenomenological caveman's phenomenal state space. As soon as the system does not need a globally available self-model, it simply tums it off. Together with the model the conscious experience of selfhood disappears. Sleep is the little brother of death.

8.2 Preliminary Answers

In the second section of chapter I I offered a small set of questions to guide us through the complex theoretical landscape associated with the phenomenon of subjective experience. I will now keep my promise and return to each one of these questions by giving brief answers to them. However, recall that the longer answers can only be found elsewhere. Let us start by taking a second look at our basic notions.

What does it mean to say of a mental state that it is conscious?

First, it is important to note that there is no one single answer to this question, but that there are now many of them. How conscious a mental state is depends on the target domain and on the degree of constraint satisfaction. Consciousness is not an all-or nothing phenomenon. There are degrees of phenomenality (for a first and simple example, think of Lewis qualia, Raffman qualia, and Metzinger qualia, as described in section 2.4). And as constraints are themselves theoretical entities, the degree of phenomenality or consciousness exhibited by a certain mental state is not only an objective property but is also relative to a given theory. Third, any answer will depend on how we choose to individuate mental states, that is, the level of description we choose to give our answer on.

If a mental state is conceived of as a representational state, something that is described as carrying a content, then this content will be minimally conscious if it is, at the same time, integrated into a virtual window of presence (an internally generated "Now") and into a single, coherent, and globally available model of reality while earlier processing stages-and therefore its representational character as such-are attentionally unavailable, that is, if it is also a transparent forro of content. The minimal degree of constraint satisfaction needed to speak about the phenomenon of "appearance," the phenomenon of consciousness at all, involves constraints 2 (presentationality), 3 (globality), and 7 (transparency). Conscious experience consists in the activation of a coherent and transparent world-model within a

window of presence. On the level of phenomenal content this is simply equivalent to "the presence of a world." Please note that such a minimal version of conscious experience is not yet subjective experience in terms of being tied to a consciously experienced first-person perspective (it is only subjective in the very weak sense of being an internal reality-model within an individual organism), and that this notion still is very simplistic (and probably empirically implausible), because it is completely undifferentiated in its representation of causality, space, and time. A system enjoying minimal consciousness as exclusively described by the conjunction of Constraints 2, 3, and 7, would be frozen in an eternal Now and the world appearing to this organism would be devoid of all internal structure.

Please note how, in more complex configurations, there may be individual states not satisfying the transparency constraint: As soon as what I have called a "world zero" is in place, phenomenal simulations become possible (see the "world-zero hypothesis" for the evolutionary function of consciousness, as proposed in section 2.3). The system may now develop phenomenal simulations, conscious mental states that are experienced as representational states. Short-terco memory and a single, integrated world-model are strictly necessary for phenomenal experience. In more complex organisms (like ourselves) transparency isn't. This is so because once conscious world-models are established, phenomenally opa que forms of content can be integrated into them.

We must therefore ask how phenomenal mental content could become stronger, by exhibiting a higher and potentially variable degree of constraint satisfaction. In principie there are now many ways to describe what I called differentiated consciousness in chapter 3. If we add a mereological internal structure in terms of constraint 4 (convolved holism), we allow for multimodal scene segmentation and the emergence of a complex situation. However, if we do not want to assume the unlikely case of one single, presegmented scene being frozen into an eternal Now on the phenomenal level, we have to add temporal structure in terms of constraint 5 (dynamicity). At this stage it is possible to have phenomenal experience as a dynamically evolving phenomenon on the level of content, to have an interrelated hierarchy of different contents that unfolds over time and possesses a dynamical structure. Differentiated consciousness, therefore, results from adding an internal context and a rich temporal structure.

The decisive step is the one leading to subjective consciousness. This is the level at which consciousness begins to approach the complexity we find on the human level of organization, and the level at which it becomes a truly theoretically interesting phenomenon. By adding constraint 6 (perspectivalness) to constraints 2, 3, 4, 5, and 7, we introduce a consciously experienced first-person perspective into phenomenal space. The space of experience is now always centered on an active self-representation. A PSM, a transparent and globally available self-model, as well as a PMIR, a transparent and globally available

model of ongoing subject-object relations, is in existence and integrated into working memory (see chapter 6 for details). Aperspectivally conscious mental state, therefore, is one whose representational content has been integrated into a phenomenal model of reality that is structured by a PMIR. There is also an alternative formulation allowing us to describe all and only those states as subjectively conscious which are currently ion the focus of experience. A truly conscious mental state would then be one that currently constitutes the object component of the phenomenal model of the intentionality relation (PMIR; see section 6.5).2

2. This also allows for a definition of what the fringe of consciousness actually is (Mangan I993). Every IR' resentational content that fulfills constraints 2,3, and 6, while not being integrated into either the subject or object component of the PMIR, constitutes the fringe of phenomenal awareness.

If we demand the satisfaction of this constraint, we pick out a much more interesting class of representational systems: the class of systems of which it can actually be said that they enjoy subjective experience in the true sense of the word. One may speculate that all those vertebrates possessing a PSM plus at least some rudimentary form of attentional processing belong to this level of constraint satisfaction. They have a conscious self (however simple it may be) and they generate a phenomenally experienced "arrow of intentionality" pointing from the attending self to various perceptual objects. They have a simple, subsymbolic PMIR. Such systems, although a subjectively experienced flow of time involving duration and change against the background of a specious present would already be available for them, would not yet have an explicit phenomenal representation of past and future, of possible worlds, and possible selves.

An even richer degree of phenomenality is constituted by cognitive subjective consciousness. If we add constraint 8 (offline activation) and if we assume a spectrum from transparent to opaque representations (see section 3.2.7), we arrive at a yet more specific class of phenomenal systems. These systems would be able to selectively engage in the activation of globally available representational structures independently of current external input, and given that these structures would exhibit a certain degree of opacity, the fact that they were now operating with representations would therefore be globally available to them and could be integrated into their self-model. In other words, such systems could not only in principle engage in future planning, enjoy explicit, episodic memories, and start genuinely cognitive processes like the mental formation of concepts, these systems could also for the first time represent themselves as representational systems, on whatever minimal a scale. They would be thinkers of thoughts. They would be like total flight simulators that have started to simulate the pilot as a simulator. Through the running of

phenomenally opaque simulations, they would be able to finally escape naive realism, previously generated by a full satisfaction of the transparency constraint on all levels of content. For such systems, the difference between reality and appearance would for the first time become available for attentional and metacognitive processing. Therefore, they would now possess the resources to develop a conception of consciousness itself, of the phenomenon of appearance as such. They could then become what I would (in alluding to Daniel Dennett's notion of a "second-order intentional system"; see, e.g., Dennett I987b, p. 243 ff.) term a "second-order phenomenal system": a being that can consciously experience the fact that it currently undergoes conscious experiences itself. It may well be that human beings are the only biological creatures on our planet fulfilling this additional condition to any interesting degree. Please also note how the adaptivity constraint (section 3.2.II) still excludes artificial systems as bearers of phenomenal mental states. I return to this issue below when giving a preliminary answer to the last question on this list.

Alternatively, what loes it mean of a conscious system-a person, a biological organism, oran artificial system-if taken as a whole, to say that it is conscious?

The transition from state consciousness to system consciousness is rather straightforward and simple. A system is conscious to the degree to which its mental states satisfy the criteria mentioned aboye. Any system possessing representational mental states, but no virtual window of presence and no single, global, integrated, and transparent model of reality, is unconscious. So even if the logical subject of predication is not a subsystemic state, but the system as a whole, ascribing phenomenality never is the same as ascribing one single, and primitive, property with the help of a one-place predicate (for which there would then not exist a noncircular definition). Ascribing phenomenality always consists in determining the degree of constraint satisfaction on multiple levels of description. Making the transition from state consciousness to system consciousness just means to exchange microlevels for macrolevels in tercos of the logical subjects and the possible predicates constituting those levels of description. There may be interesting and highly relevant constraints, which can be exclusively discovered and applied on the whole-system level only-for instance, when investigating the social correlates of complex forros of phenomenal experience (see section 6.3.3). In particular, if the macrolevel is not simply the whole-system level but the personal level of description, a fundamental transition to an entirely new dimension is made. This may constitute a second fundamental distinction between human beings and other conscious beings on our planet. To give an example, in conscious systems, which, by accepting certain normative standards (epistemically justified or not), have begun to phenomenally experience themselves and others as rational individuals and as moral subjects, we have to explain not only the phenomenal experience of "selfhood" but

also that of "personhood." This brings about a whole new set of properties and predicates on the whole-system level.

In terms of individuating characteristics for mental states, it is interesting to note that there could conceivably be afunctional phenomenal states, which are not representational states at all (e.g., in dreams or some kinds of hallucinations). Such states could contribute to conscious experience, while not representing anything for the organism as a whole. According to our teleofunctionalist background assumption they would have phenomenal, but not intentional content. In this case they will have to be individuated on a lower leve] of description, for example, as purely functional states currently functionally integrated into the mechanism that creates the organism's experiential present and its world-model. In this case, it would be their causal role that has been integrated, but not their represen tational contents. Cal] this "vehicle consciousness."

What does it mean to say of a mental state that it is a part of a given system 's self-consciousness?

All mental states constituting phenomenal self-consciousness are characterized by a further content property, the property of mineness. Mineness represents ownership on a nonconceptual level (see section 6.I). In conscious processing, mineness creates a prereflexive, and fully transparent sense of ownership. It is a property of a particular form of phenomenal content that, in our own case, is accessible on the level of inner attention as well as on the level of self-directed cognition. It is available to introspection3 and introspectior>4 (see section 2.2). In pathological situations, the distribution of this property across phonomenal space can vary considerably. A mental state is part of a given system's seifconsciousness if it has been integrated into the system's PSM (see chapter 6). Its representational content has then become a part of the system's phenomenal self. Functionally, any system property currently represented in the PSM is an appropriated property. If, in unusual configurations (see chapter 7), a representational state satisfying the constraints for phenomenality cannot be integrated into the PSM, it automatically becomes a part of the world-model and its content is now experienced as external. For instance, a conscious thought could not be phenomenally owned any more, if-as in some cases of schizophrenia-the system is unable to embed it in its PSM. It would then not be my thought anymore. Or a body part, as in unilateral hemineglect, could drop out of the phenomenal self, if the system is for some reason unable to integrate it into the globally available partition of its self-model. Phenomenally, it would then not be my own body part anymore.

What does it mean for any conscious system to possess a phenomenal self? Is selfless consciousness possible?

First, it is important to understand the central ontological claim: No such things as selves exist in the world. All that exists are certain information-processing systems meeting the constraints for phenomenality while operating under a transparent self-model. At least for all conscious beings so far known to us, it is true that they neither have nor are a self. Biological organisms exist, but an organism is not a self. Some organisms possess conscious self-models, but such self-models certainly are not selves-they are only complex brain states. However, if an organism operates under a phenomenally transparent selfmodel, then it possesses a phenomenal self. The phenomenal property of selfhood as such is a representational construct; it truly is a phenomenal property in tercos of being an appearance only. For all scientific and philosophical purposes, the notion of a self-as a theoretical entity-can be safely eliminated. What we have been calling "the" self in the past is not a substance, an unchangeable essence, or a thing (Le., an "individual" in the sense of philosophical metaphysics), but a very special kind of representational content: the content of a phenomenally transparent system-model (see section 6.2). It is the content of a self-model that cannot be recognized as a model by the system using it. The phenomenal experiences of substantiality (ie., of being an independent entity that could in principie exist all by itself), of having an essence (ie., of being defined by possessing an unchangeable innermost core, an invariant set of properties), and of individuality (ie., of being an entity that is unique and cannot be divided) are special forros of conscious, representational content as well. Possessing them was evolutionary advantageous, but as such they are not epistemically justified. As such, they are not a form of knowledge, aithough they play an important functional role.

On the functional level of description, a phenomenal self, again, is not a substance or an individual-be it physical or nonphysical-but an ongoing process: the process of self-modeling, as currently integrated into working memory and the organism's globally available world-model. This process can be interestingly described as a process of self-containing, of functionally achieving ownership for a subset of the system's causal capacities. Self-modeling is causal self-appropriation. What we called the "phenomenal self-shadow" earlier is determined exclusively by the machinery of internal functionai properties. On the neurobiological level, the phenomenal content of the self-model supervenes locally. This means that in biological organisms, every phenomenal self possesses a minimally sufficient neural correlate. Given this correlate, a conscious self will come into existence by nomological necessity.

A phenomenal self appears if a certain property is instantiated, the phenomenal property of selfhood. In its core, this property is a representational property. Interestingly, it is

brought about by a special form of epistemic darkness, by a lack of introspectively available information. It is important to note this point: phenomenal selfhood results from phenomenal transparency, but from epistemic opacity. According to the SMT, phenomenal selfhood is a lack of introspective self-knowledge. I have called this structural characteristic of the neurophenomenological caveman's conscious mind "autoepistemic closure" (see sections 2.3, 3.2.7, and 6.2.6), referring to it as an "inbuilt blind spot," a structurally anchored deficit in the capacity to gain knowledge about oneself. It is important to understand that autoepistemic closure as used in this book does not refer to cognitive closure (McGinn I989b, I99I) or epistemic "boundedness" (e.g., Fodor I983, p. I20) in terms of the perhaps principled unavailability of theoretical, propositionally structured selfknowledge. Rather, it refers to a closure or boundedness of attentional processing witb regard to one's own internal representational dynamics. It is a limitation in mental resource allocation expressed on the level of nonconceptual content. Autoepistemic closure, in the current context, consists in human beings in ordinary waking states, using their interna¡ representational resources-that is, by introspectively guiding attention-not beii g able to attentionally penetrate into earlier processing stages in the ongoing construction of their conscious self-modei. Of course, there may be good evolutionary reasons for this: Attentional availability uses precious computational resources, and a transparent seli model-a realistic self-model-has the functional advantage of making its be maximally egotistic.

Is selfless consciousness possible? All consciousness is selfless, in that a self is not represented in it, but only a physical, representational system-but transparently, in the mode of naive realism, as it were. Because the PSM is transparent, the system constantly operates under the condition of what I have called a naive-realistic self-misunderstanding (see section 6.2.6). Metaphorically speaking, it confuses itself with the content of its own PSM. Just as with color qualia there is nothing in the extemal world that nicely and systematically maps on the chromatic primitives of conscious color vision, so there is no single entity in or outside the system that directly corresponds to the primitive, prereflexive feeling of conscious selfhood. In principie there are two ways in which a phenomenal system could lack this feeling, in which selfless consciousness is conceivable within the present framework.

First, it is possible for a system to satisfy all other constraints for consciousness, without having a self-model. It could have a world-model, but no self-model. Probably many simple organisms on our planet belong to this phenomenal system class. If the system at least satisfies constraints 2, 3, and 7, but without possessing a centered model of reality, then it will instantiate selfless consciousness. Such organisms may have unconscious proto-selves, for example, in terms of the elementary form of functional selfappropriation that comes with homeostasis and rudimentary emotions, but no distinct conscious representation

directed at the intentional object of the organism as a whole. There would be the appearance of a world, but no one to appear as currently being directed toward this world. Phenomenologically, the light would be on, but no one would be at home. There would be no explicit little red arrow, and only a flight simulator, but no total flight simulator.

There is, however, a second possibility and it is of much greater philosophical interest. In section 6.4.2 we saw that the human self-model is interestingly characterized by exhibiting a continuum ranging from full transparency to opacity, typically ascending from the sensory aspects of bodily self-awareness to purely cognitive levels of self-reference and reflexive self-consciousness. Try to imagine a PSM that was fully opaque. Imagine a system that-all other aspects held constant-is characterized by the fact that constraint 7, the transparency constraint, is not satisfied for its self-model at all. Earlier processing stages would be attentionally available for all partitions of its conscious selfrepresentation; it would continuously recognize it as a representational construct, as an internally generated internal structure. The SMT makes the following prediction: Phenomenologically, this system would not have a self, but only a system-model. It would not instantiate selfhood. Functionally, it would still possess all the computational and informational advantages associated with having a coherent self-modei, at the price, however, of a somewhat higher computational load. In addition, it would have to find a new solution to the problem of not getting paralyzed by an infinite loop of self-representation, to the problem of avoiding an infinite regression in the absence of transparent primitives. But possibly it could still operate under a centered model of reality, even if this model were not phenomenologically centered anymore. What the neurobiological characteristics of such a system would be is presently unclear. However, it may be interesting to note a specific phenomenological analogy. There is one type of global opacity that we discussed in our last neurophenomenological case study, namely, the lucid dream (see section 7.2.5). In the lucid dream the dreamer is fully aware that whatever she experiences is just the content of a global simulation, a representational construct. It is also plausible to assume that there are state classes in the phenomenology of spiritual or religious experience resembling this configuration-but only during the waking state. Now imagine a situation in which the lucid dreamer would also phenomenally recognize herself as being a dream character, a simulated self, a representational fiction, a situation in which the dreaming system, as it were, became lucid to itself. This is the second possibility for selfless consciousness under the theoretical framework proposed here. I am, of course, well aware that this second conception of selflessness directly corresponds to a classical philosophical notion, well-developed in Asian philosophy at least 2500 years ago, namely, the Buddhist conception of "enlightenment." However, let us adopt a metaphysically neutral terminology here and call this phenomenological state class "system consciousness." A representational system has

system consciousness if and only if it operates under a phenomenally opaque system-model, but not under a self-model.

What the first possibility and the second possibility have in common is that they are logical possibilities; they can be coherently described and conceived of. Whether they are nomologically possible neurophenomenological configurations is an open question. For instante, there could be fundamental neurocomputational reasons that make such selfless modess of reality at least highly unlikely, hard to sustain, or generally unstable. Assuming the second case, it may turn out that any representational system needs some kind of transparent primitive, and that this is true for human self-consciousness in particular. On the other hand, please note that all that is needed for generalized opacity is the availability of earlier processing stages for introspective attention, but not a permantly realized form of actually ongoing access. For the first class of phenomenal systems, it is plausible to assume that many lower animals on our planet function in this way. Autophenomenological reports given by human beings about selfless states of the second type, however, will typically not impress philosophers much, because they contain an inherent logical fallacy: How can you coherently report about a selfless state of consciousness from your own, autobiographical memory? How could this episode ever constitute an element of your own mental life? Such reports generate a performative self-contradiction, because you deny something that is presupposed by what you are currently doing. (For a more mundane example; "I am probably the most modest person I have ever met.")

In any case, it is interesting to note a second common characteristic of the first and second selfless configurations: they are phenomenally impossible, and therefore extremely. counterintuitive. In section 2.3, I introduced the notion of phenomenal possibility, as a property of all states of affairs or worlds which, as a contingent matter of fact, we can actually consciously imagine or conceive of-all those states of affairs or worlds Wich, can enter into conscious thought experiments, into cognitive operations, or explicit planning processes. We also saw that what is phenomenally possible is always relative to a certain class of concrete conscious systems, to their specific functional profile, and to the deep representational structure underlying their specific form of phenomenal experience. For beings like us, the goal of deliberately simulating a noncentered, selfless reality is strictly incompatible with our representational architecture. We cannot truly imagine the world as viewed from nowhere, pace Nagel. When, earlier, I asked the reader to imagine a fully opaque PSM or discovering that oneself is a dream character, I was asking for something impossible. Children discover this impossibility for the first time when trying to imagine how the world will be after they are dead. Adults certainly can phenomenally simulate noncentered worlds within a centered world zero, but there will always be a phenomenal self experienced as doing the imagining. The view from nowhere always is your view-or it could not be an element of your autobiographical memory about which you could later

report. In short, the self-model theory of subjectivity is a theory which, even if strongly supported by good arguments and empirical data, will always remain counterintuitive. Even if you are intellectually convinced by the current theory, it will never be something that you can believe in.

What does it mean to say of a mental state that it is a subjective state? Is nonsubjective consciousness possible?

First, it is important to note that so far we are only talking about phenomenal subjectivity, that is, of subjectivity as phenomenally experienced. There is a more trivial reading of subjectivity (previously introduced as "functional subjectivity"; see section 2.2), amounting to the fact that information has been integrated into an exclusively internal model of reality, active within an individual system, and, therefore, giving this particular system a lund of privileged introspective access to this information in terms of uniquely direct causal links between this information and higher-order attentional or cognitive processes operating on it. If this internal model of reality satisfies the minimal constraints for perspectival phenomenality, then three major interpretations of "phenomenal subjectivity" result.

First, I experience everything as subjective that is an element of my conscious world-model. Even, if I don't experience it as mental I learn (e.g., through visual illusions and other cases of sensory misrepresentation) that, strictly speaking, my world is only my world and that others may have a different kind of phenomenal experience. To be sure, my world-model remains transparent, but, through experience, the fact that in all its reliability it nevertheless must be a model becomes cognitively available to me. And this event changes my PSM: I am now someone who consciously experiences himself as knowing this very fact. This is a weak, cognitively mediated form of phenomenal subjectivity from the firstperson point of view. There is also a straightforward third-person reading of this first notion of phenomenal subjectivity: any system that has a conscious world-model has phenomenally subjective states. Please note how cognitive subjéctivity emerges from an interna) simulation of just this third-person reading: Cognitive subjectivity results when a system representationally distances itself from its own world zero.

The second interpretation of "phenomenal subjectivity" is more interesting. Any representational content that has been integrated into a PSM is phenomenally subjective. Whatever is represented under a PSM is an element of an individual system's self-consciousness. It is now phenomenally owned, by gaining the additional phenomenal property of "mineness." Phenomenal selfhood creates internality in the sense that something is portrayed as currently belonging to the center of representational space, as being a property of this subject. To be a subjective content then means to be a state of the phe-

nomenal self, to be seamlessly integrated into it. However, in order to properly understand what has just been said, we need to understand how a phenomenal self can be portrayed not only as an ego but as a subject-as a subject of knowledge, an autonomous agent, or, to take the most simple case, as a currently attending self. Are there neurophenomenological configurations in which a phenomenal self is in existence, but no conscious subject? Is it possible for a system to have a conscious world-model plus a conscious self-model but no phenomenally subjective states?

What turns a phenomenal self into a conscious subject is the fact that it is transiently integrated into a yet more comprehensive kind of globally available representational structure: the PMIR. Phenomenal subjectivity in a truly interesting sense only emerges at this stage. It is the moment in which the system experiences itself as directed at a possible object of knowledge, an action goal, or a perceptual object. Truly subjective states are those that are integrated into the representation of a specific relation, namely, a self-object relation. I explained the notion of a PMIR at length in section 6.5, and there I also, in quoting from the work of Antonio Damasio (Damasio I999, p. 263; for a brief case study, see p. I0I ff.), pointed out how akinetic mutism may be a particularly circumscribed and salient example of a rare neurophenomenological configuration, the possibility of which is predicted by the SMT. Bilateral anterior damage to the cingulate gyrus and bilateral media) parietal damage lead to a situation which can be described as, first, the absence of a PMIR, while, second, a coherent conscious model of the world functionally centered by a phenomenal self is retained. Full-blown conscious experience-phenomenal subjectivity in a philosophically interesting sense-is more than the existence of a conscious self, and it is much more than the mere presence of a world. It results from the dynamic interplay between this self and the world, as situated in a lived, embodied present. In the patient with akinetic mutism we arguably have a situation in which there is a PSM, but no PMIR. The patient is awake, but not a subject. He may stare vacantly at the world or orient mechanically toward some visual object, but he never is a self in the act of seeing. Tb represent the act of seeing you need a PMIR. The patient is phenomenally embodied, but not present, because he is not phenomenally situated-situatedness is precisely what is established through the ongoing, dynamic construction of a PMIR. I have, on philosophical grounds, introduced the PMIR as a distinct theoretical entity on the phenomenological, representational, and functional levels of description (see chapter 6). I am therefore committed to the empirical prediction that there will be a distinct neural correlate as well. In fact, candidates for the necessary components of the neural correlate of this specific kind of a PMIR are already under discussion (e.g., the cingulate gyrus, certain thalamic nuclei, and the superior colliculi; cf. Damasio I999, p. 260ff.).

Returning to the level of philosophical analysis, I propose to treat the notion of phenomenal subjectivity as exactly that which may be absent in akinetic mutism. I argue that it

is this kind of phenomenal content-a transient, dynamic integration of subject and object-that many of us intuitively regard as the essence of conscious experience. The most interesting sense of phenomenal subjectivity is the one that comes with constraint 6, the perspectivalness constraint. A truly subjective representational content is one that is an element of a perspectiva) model of reality, one that is structurally dominated by a PMIR.

Again, there is an alternative and more narrow formulation allowing us to describe all and only those states as subjectively conscious which are currently in the focus of experience. A truly conscious mental state would then be one that currently constitutes the object component of the PMIR. Viewed as a forro of representational content, its subjectivity then consists in being explicitly linked to a phenomenal self-in its contribution to a more comprehensive mental structure, a relational representation of the act of experience. To give an example, as empirical research on change blindness (Mack and Rock I998) shows for visual consciousness, there clearly exists an attentional PMIR in the visual domain: its object component is simply what we experience in an integrated fashion. Once attention is released, visual objects dissolve back into "proto-objects" and all informational content is lost. The phenomenon of change blindness demonstrates how systems like ourselves only integrate what becomes the visual object component of the PMIR, thereby minimizing the computational load on our brains. Computationally speaking, it is not necessary to keep all objects represented simultaneously, all that is needed is the capacity to access object identity whenever necessary. There is also, of course, nonattentional extraction of scene structure, because attention is not the central gateway through which all conscious information must pass; the attentional bottleneck applies only to coherent objects. Preattentive vision gives us scene structure, which is everything that can be seen before the arrival of the limited-capacity selection mechanism on the object. In this sense, and according to the third possible reading of "phenomenal subjectivity," preattentive scene structure would be a nonsubjective form of conscious content. This third interpretation of phenomenal subjectivity not only follows the philosophical intuition of "subjectivity-as-focal-representation-only" but nicely demonstrates what it means to say that conscious experience truly is a graded phenomenon.

What does it mean to speak of whole systems as "subjects of experience?"

Again, the transition from the subjectivity of states to the subjectivity of systems is rather straightforward and simple. A system is subjectively conscious to the degree to which its II mental states satisfy the constraints mentioned aboye. Based on the second interpretation !i of "phenomenal subjectivity" we can now say that any system possessing a virtual window of presence and a single, globally integrated, and transparent model of reality but no PSM and no PMIR is not a subject of experience. Generally speaking, to

become a true subject of experience you have to represent the world under a stable PMIR. However, there are borderline cases, like the nonlucid dreamer who possesses a highly unstable PMIR or the patient who suffers from akinetic mutism. Although such a patient has minimal self awareness and no perspectiva) form of first-person experience, I would plead that all systems-human or not-belonging to this phenomenological class be treated as genuine subjects of experience. Why?

The notion of a "subject of phenomenal experience" is of great relevance not only for philosophy of mind but for ethics as well. Without going into any technical issues here, I would argue that everything that is capable of conscious suffering should automatically be treated as a moral object. Put simply, a moral object is something that belongs to the domain of things with regard to which our actions should be morally justifiable. Call this the "principle of negative utilitarianism": Whatever else our ethical commitments and specific constraints are, we can and should certainly all agree that, in principle, the overall amount of conscious suffering in all beings capable of conscious suffering should be minimized. This seems to be a simple principle of solidarity among all conscious creatures that are mortal, and able to feel physical pain or to suffer emotionally, intellectually, or otherwise. Whatever is a phenomenal subject of experience should immediately be treated as a moral object. It is interesting to note how the SMT predicts that many animals on this planet (as well as the first artificial subjects of exPerience that may one day evolve; see Metzinger 200I) are phenomenal subjects-but not yet moral subjects. They cannot mentally represent norms and are in principle unable to impose moral obligations unto themselves. Although they have no conscious first-person perspective, although they have no cognitive, let alone moral, first-person perspectiv, they should definitely be treated as moral objects. It is important to note the simple facá that all of the aboye does not imply that they cannot suffer. Maybe suffering is evef more intense in simpler creatures that don't have the mental capacities to cognitively di tance themselves from their pain or understand the potential meaning their suffering might have.

Remember the patient with akinetic mutism. Arguably, he is not capable of first-person, perspectiva) suffering, because he has no phenomenal first-person perspective. He cannot represent reality under a stable PMIR. However, he can certainly own physical pain, which, for instance, might occur in his body. He has rudimentary self-awareness. I would argue that even phenomenal ownership alone is enough for suffering: We should treat every representational system that is able to activate a PSM, however rudimentary, as a moral object, because it can in principie own its suffering, physical or otherwise. It is the phenomenal property of "mineness," the phenomenal, nonconceptual sense of ownership, which counts for ethical purposes. Without phenomenal ownership, suffering is not possible. With ownership, the capacity for conscious suffering starts to evolve. We would never deliberately hurt a patient with akinetic mutism, even if he could neither talk nor

move and even if all we could elicit is the well-known vacant stare. The same principie should hold for all other weakly conscious systems, for ail creatures characterized by low degrees of constraint satisfaction. In particular, we should take care to always stay on the safe side: As soon as there is evidence that something is a weak phenomenal subject of experience, as soon as there are indicators for the existence of a PSM, we should automatically treat it as a moral object. Of course, much more needs to be said about negative utilitarianism, its potentially limiting principies, and about the connection between philosophy of mind and ethics in general. And, of course, it is obvious how cognitive neuroscience now starts to gain increasing relevance for ethical issues. As a scientific discipline, it has the great potential to make extremely valuable contributions in the future in tercos of precisely pinning down objective indicators for the existence of a PSM in a given nervous system, in empirically defining inclusion criteria for the class of phenomenal subjects, and thereby for the class of moral objects. But this is not the place for this type of investigation.

What is a phenomenal first-person perspective, as opposed, for example, to a linguistic, cognitive, or epistemic first-person perspective?

A linguistic first-person perspective appears with the mastery of the first-person pronoun "l." For a cognitive first-person perspective to emerge it is not only necessary to have thoughts that can be expressed using "I." What is necessary is the possession of a concept of oneself as the thinker of these thoughts, as the owner of a subjective point of view. An epistemic first-person perspective comes into existence if the system's model of reality, as structured through a PMIR, is not only characterized by its phenomenal content but also as possessing intentional content. It is then described as a structure that not only mediates conscious experience but also knowledge. It is interesting to see how the phenomenal firstperson perspective is a necessary foundation for ali the richer, more complex forms of subjectivity just mentioned, and how it is at the same time fully autonomous. Every philosophical investigation of higher-order forms of subjectivity-be they mediated through linguistic and cognitive self-reference, through propositional forms of structured self-knowledge, or even through social interactions-will inevitably have to rest on a convincing account of the PMIR. Let us have a short look in four brief steps.

First, what is a phenomenal first-person perspective? And what does it mean that it is autonomous? A phenomenal first-person perspective is realized by any system poocessing a transparent PSM plus a transparent PMIR. In particular, every system satisfying constraints 2, 3, 6, and 7 will have a phenomenal first-person perspective. More realistically,' it is important to note that all candidates for phenomenally experienced perspectivalness actually existing in the part of the universe known to us are highly likely to satisfy all the

constraints developed in chapter 3 with the exception of constraint 8, the capacity for offline activation. Of course, given the terminological machinery developed in chapters 2, 3, 5, and 6, it is now possible to offer many fine-grained descriptions of different grades of first-person phenomena, of different degrees of constraint satisfaction-like consciousness in general, selfhood or phenomenally experienced perspectivalness is not an all-or-nothing phenomenon. How perspectival a mental state is depends on the target domain and on the degree of constraint satisfaction, and any judgment is theory-relative. However, I will not enter into a discussion here, but highlight just one single aspect: It is empirically plausible to assume that a large majority of phenomenal systems currently known to us will have only very limited resources to run consciously experienced mental simulations and self-simulations (see sections 2.3 and 5.3.) They will have dynamic and somewhat convolved phenomenal models of reality, including a rudimentary self-mo4el and a simple attentional first-person perspective (see sections 6.4.3 and 6.5.2). But, put simply, they will not be thinkers of thoughts and will have only limited capacity for explicit episodic memory and future planning. In particular, many of them will lack an opa que partition of their self-model (see section 6.4.2).

The autonomy of the phenomenal first-person perspective consists in that it can exist in nonlinguistic creatures and that it does not presuppose strong cognitive first-person phenomena in any way. A fully transparent PMIR is enough. You do not peed to have a concept of yourself as operating under a phenomenal first-person perspective in order to possess it, neither linguistic nor mental. On the contrary, all empirical indicators strongly point to the hypothesis that abstract forros of self-representation evolved out of and are anchored in subsymbolic (e.g., spatial, proprioceptive, motor, and emotional) forms of Self representation, that any conceptual point of view can only be acquired via a nonconceptual point of view (see Bermúdez I998; Metzinger I993; and chapter 7). To establish what called the phenomenal presence of a knowing self in section 6.5.2 (see also the intimmately related notions of a "juxtaposition of self and object" and of a "self in the act of knowing" in the work of Antonio and Hanna Damasio; Damasio and Damasio I996a, p. I72; I996b, p. 24 ff.; Damasio I999, p. I68 ff.), it fully suffices that the PMIR is constituted by attentional, that is, subsymbolic, mechanisms. Call this "subdoxastic subjectivity." Cognitive processing and concept formation are not needed to activate a PMIR. Attentional subjectivity (see section 6.4.3) is already a full-blown first-person phenomenon.

Third, the autonomy of the phenomenal first-person perspective also consists in not presupposing an epistemic first-person perspective. Please recall that all this time we have been discussing phenomenal content only. Phenomenal content supervenes locally. It follows that even the highest and most complex form of phenomenal content that human beings are arguably capable of, including all its higher-order variants emerging through reflexive self-consciousness and social cognition, is fully determined by the properties of its

minimally sufficient neural correlate. Isolated portions of brain in a vat could generate a PMIR. What they could never generate is first-person knowledge. A minimally sufficient neural correlate in a vat could not even know what kind of properties the current PSM supervenes on, because, apart from the fact that it could hardly count as an epistemic subject, it would lack independent means of verification.

What is a linguistic first-person perspective? The "principle of phenomenal reference" introduced in chapter 2 states that one can only deliberately speak and think about those things that one also consciously experiences. Only phenomenally represented information can become the object of linguistic or cognitive reference, thereby entering into communicative and thought processes that have been voluntarily initiated. It is important to fully understand this principle. If you want to linguistically refer to, say, Gódel's theorem or to a friend living on the other side of the earth, you can only do so if you have, in whatever sketchy and rudimentary way, phenomenally simulated them. There must be a representation of them that is globally available for speech control and cognitive processing.

Linguistic reference functions via phenomenal representation. Talking in your sleep or during light anesthesia or an epileptic automatism is not linguistic reference at all, because it is not agency, it is automatic motor behavior blindly producing speech output, without this output having been voluntarily initiated. A speech act always presupposes a phenomenal first-person perspective. The same is true of thought. Only phenomenally represented information can become the object of explicit cognitive reference, thereby entering into further thought processes which have been voluntarily initiated. If you refer linguistically to events in the distant past or future you can only do so by first representing them within your own virtual window of presence. If only very briefly, they have to become an element of global working memory.

There is a related principle for linguistic and cognitive self-reference. Not only reference de re, but referente de se has to be internally modeled while it is taking place. SMT proposes that the PSM is the neurocomputational tool making this possible. In short, what is needed for stronger forms of subjectivity is not only reference from the first-person point of view but the capacity of mentally "ascribing" this act of reference to oneself while it is taking place. However, it is empirically more plausible that this "ascribing" take$ place in a dynamic, subsymbolic medium and in an ongoing fashion, like a permanent ("transcendental") process operating in the background. We have to keep this in mind when using the concept of "cognitive self-reference": We are not talking about discrete symbol tokens, but about dynamical self-organization in human brains. Cognitive self-reference always is reference to the phenomenal content of a transparent self-model. More precisely, it is a second-order variant of phenomenal self-modeling, which, however, is mediated by one and the same integrated vehicle of representation. The vehicle is not a thing, but a process. The capacity to conceive of oneself as oneself* consists in being able to activaté a

dynamic, "hybrid" self-model. Phenomenally opaque, quasi-symbolic, and second-order representations of a preexisting phenomenally transparent self-model are being activated and continuously reembedded in it.

Recall the discussion in section 6.4.4. Weak first-person phenomena are those in which, for instance, animals can be conceived of as operating under an egocentric world-model forming the center of their own universe and the origin of their own perspective. Such simpler animals do not have a hybrid self-model, because they generate no opaque states that they could continuously reintegrate into it. Using Lynne Baker's terminology we can now say that all sentient beings are conscious subjects of experience, but not all of them have first-person concepts of themselves. For Baker, only those who do are fully self-conscious in an interesting sense (Baker I998, p. 328; see also note I8 in chapter 6, p. 396). Under the SMT simple sentient beings would use an integrated, global, and transparent model of the world functionally centered by a transparent self-model to regulate their own behavior. In Bakerian terminology, such organisms could be said to be solving problems by employing perspectival attitudes, while not yet possessing a concept of themselves as a subject. First-person phenomena in a stronger and more interesting sense, however, are not only characterized by the necessary condition of possessing a self-world boundary and being able to differentiate between the first and thid person but include the capacity to possess this distinction on a conceptual level, and the act of currently using it. In the terminology so far introduced, this means that the existente of a preattentive self-world boundary and the difference between first- and third-person attributions are cognitively available in terms of introspection$_{2/4}$. It is not only necessary to have thoughts that can be expressed using "I." What is necessary is the possession of a concept of oneself as the thinker of these thoughts, as the owner of a subjective point of view. In short, what is needed is not only reference from the first-person point of view but the capacity of mentally "ascribing" this act of reference to oneself while it is taking place. The PSM of human beings enables this important step by possessing a transparent and a stable opaque partition at the same time.

Last, we have to ask, What is an epistemic first-person perspective? Here is my answer: Epistemic perspectivalness comes about if a fact is correctly represented under a PMIR. If some intentional content is integrated into a PMIR, or, more precisely, if it constitutes its object component, then it is a perspectival form of representational content. First-person knowledge is knowledge under a PMIR. And this finally tells us what causes the core problem in the philosophy of consciousness, the epistemic asymmetry (Jackson I982): If Mary-who is one of the most recent descendants of Plato's captives in the cave-finally leaves her achromatic prison and sees the blue sky and a red apple on a tree for the first time, she represents a physical fact already previously known to her. But now, for the first time, this fact is integrated finto the object component of her PMIR, for the first time

she represents this aspect of reality under a transparent PMIR. She generates a new epistemic possibility, by gaining a new mode of knowledge. The new mode of presentation is being known under a PMIR. The same physical fact-that the neural correlate for certain conscious color experiences is currently active-is now for the first time represented to her as something she is directed to, under a PMIR. Moreover, it is represented transparently: Mary has no noncognitive introspective3 access to the additional fact that all that is currently going on while she sees the blue sky or a red apple is a representational process. Even after Mary leaves her prison, she is still a neurophenomenological cavewoman. She is only a very distant relative of Plato's prisoners.

A rational theory of consciousness will have two major explanatory goals. First, How can full-blown, perspectival phenomenal experience be ontologically reduced? If there are principled obstacles, how can these obstacles be described so precisely that these descriptions in themselves constitute a growth of knowledge? Second, How can we at the same time give a plausible account of the fact that it is epistemically irreducible (see Walde 2002)? The concept of a PMIR as introduced in section 6.5 now enables us to give a clear and straightforward answer to the second question. Phenomenal content is epistemically irreducible, because-in standard situations-it is integrated into a global model of reality structured by a PMIR. The special and hitherto somewhat mysterious fact that the phenomenal character of conscious states seems to constitute an irreducible first-person form of content can be reduced to the fact that this character is typically represented under a PMIR. And this way of gaining knowledge about your own mental state certainly is irreducible to, say, any scientific procedure producing knowledge about its neurofunctional correlate. It is another way of gaining knowledge-one that existed long before philosophy and science carne into being. On the contrary, arguably, it was the existence of a stable PMIR that made cognitive subjectivity (see aboye) and theoretical intersubjectivity possible in the first place.

So much for our first set of questions. Now we face-a number of more general questions concerning ontological, logical or semantic, and epistemological issues. They do not form the focus of this investigation, but nevertheless are of great relevance.

Is the notion of a "subject" logically primitive? Does its existente have to be assumed a priori? Ontologically speaking: Does what we refer to by "subject" belong to the basic constituents of reality, or is it a theoretical entity that could in principie be eliminated in the course of scientific progress?

Let us first limit the scope of our answer to the two notions relevant to our context: the concept of a "subject of experience" and the notion of a "phenomenal subject." A subject of experience is any system that has phenomenal states satisfying constraint 6. As soon as a system possesses not only minimal and differentiated consciousness (see section 3.2.II) but also structures its global phenomenal state with the help of a single, coherent, and temporally stable PMIR (see section 6.5), it is a subject of experience. Subjects of

experience are systems representing reality under a transparent PMIR. Therefore, there is no prima facie reason to believe that they form a category of irreducible and ontologically distinct entities. There is no intemal homunculus on any level of description: being a subject of experience is never a property of the self-model, but always a property of the system as a whole. The class of subjects of experience is formed by all systems sati ing constraints 2, 3, 6, and 7. It is a class of functional or representational architecturés. However, the truly interesting or intended class of systems, in the context of the current theory, is only formed by the maximal degree of constraint satisfaction, as explainett in sections 3.2 and 6.2. In order to have a PMIR, you need to have a transparent self-modal forming its subject component. Therefore, any system qualifying as a subject of e ente will also have a consciously experienced self. But what is the difference between a subject and a self? This question leads us to the second relevant notion.

A phenomenal subject is a specific kind of phenomenal self: a representation of the system as a whole as currently being a subject of experience, as currently being an ag bodily, attentional, or cognitive (see sections 6.4.3, 6.4.4, and 6.4.5). If we want to t phenomenology seriously, we have to conceptually integrate two additional but im constraints: you can have a phenomenal self without being a subject of experience you can be a subject of experience without cognitively knowing this fact. As expl aboye, there are many situations in which a transparent self-model is active, but currently integrated with any object component. This may not only be the case in logical configurations like akinetic mutism, but also in some everyday situations imagine states of complete exhaustion or prostration, in which you are merely v staring at the world, without truly seeing anything, without attending, thinking, or a at all; or brief transitory phases of waking up from deep sleep). Second, operating preliminary Answers a purely attentional PMIR-as many animals probably do-does not include cognitive availability of this fact, or strong first-person phenomena in Lynne Baker's sense (see section 6.4.4). You can be a subject of experience for all of your life without knowing this fact. For example, this will be true of all systems satisfying constraint 6, but not constraint 8.

From this it follows that the notion of a "subject of experience" is not logically primitive. First, there is not one simple set of syntactic or semantic rules governing the use of this expression, but according to the different degrees of constraint satisfaction there are many different ways of using the expression "subject of experience." There are even borderline cases (recall the notion of "system consciousness" introduced aboye), in which we can conceive of selfless subjects of experience, namely, all systems possessing a fully opaque subject component in their PMIR. In particular, when speaking about consciousness and phenomenal selfhood, there is no a priori implication of experiential subjectivity. As general phenomenological observations and the case studies in chapter 7 show, conscious systems not possessing an integrated phenomenal self are not only logical possibilities but

exist in the actual world. Not every conscious system has a phenomenal self. And systems possessing a phenomenal self do not necessary have to be subjects of experience or experience themselves as such. Not every phenomenal self is a phenomenal subject.

Moving on from logical to metaphysical considerations, it certainly is not necessary to assume a simple, basic constituent of reality corresponding to any of our folkpsychological or folk-philosophical notions of a phenomenal self or subject of experience. No such things as selves or subjects of experience exist in the world. What exist are natural systems operating under transparent PSMs and PMIRs, with both of these representational structures coming in many different strengths and with a long evolutionary history. We can therefore greatly simplify the ontological set of background assumptions necessary to do proper scientific psychology and cognitive neuroscience. All that exists are phenomenal selves, as instantiated by transparent self-models. For methodological purposes, no stronger assumption is necessary. The same is true of subjects of experience. We can be parsimonious by doing without the assumption that there are any basic, independent constituents of reality in this sense. For the cognitive neuroscience of consciousness and scientific psychology in general, all that exists are phenomenal models of the intentionality relation. Subjectivity is not a thing, but a property of complex representational processes unfolding in certain physical systems. In principle-and one may certainly have doubts that this would be rational in all contexts-the corresponding theoretical entities can be eliminated, and substituted by successor concepts on representationalist and functionalist leeels of description. As a matter of fact, we have already taken the first steps: A "subject of experience" is a conscious representational system satisfying constraint 6. A "phenomenal self' is constituted by the representational content oI a PSM. A "phenomenal subject" is a PSM integrated into a PMIR.

It is clearly outside the scope of my approach to develop a more detailed semantics for the indexical expression "I." Therefore, let me just very briefly sketch how we could arrive at a deeper understanding of sentences in which the word "I" is used in the autophenomenological self-ascription of experiential properties (as in "l am feeling a toothache right now").

What are the truth conditions for sentences of this type?

Self-ascriptions of phenomenal properties refer to the currently conscious content of the self-model. You can only linguistically self-refer to properties of yourself that have before been made globally available through the conscious part of your self-model. A PSM is a necessary, but not a sufficient condition for self-reference and phenomenological attitudes de se. Many different properties of oneself-social, physical, functional-can be made available for self-report through the conscious self-model. They then form its

representational or intentional content. In the special case of autophenomenological reports about the experiential character of certain aspects of self-consciousness we refer to the current phenomenal content of the self-model. Because the self-model is transparent, most people never distinguish the two situations. Intuitively, most of us therefore treat all kinds of self-reference as direct self-reference, because we normally experience them as such. Phenomenal immediacy, however, is not referential immediacy. Of course it is true that only the experiencing subject can refer to its own phenomenal states as phenomenal states. But as we now know, no kind of self-reference is ever truly direct, because it is inevitably mediated through the self-model, because it crucially depends on the subpersonal self-organization of the relevant construction mechanisms, which are introspectively invisible, that is, transparent to us. This is also true of the special case of self-related phenomenal content.

The phenomenal content of self-consciousness supervenes locally. It will have a distinct neural correlate in every single case. From a third-person perspective we can, therefore, in principie, assess the truth of an autophenomenological report by verifying the existence of its minimal sufficient physical correlate. What makes an autophenomenological statement true is the fact that its minimally sufficient correlate was functionally active at the time it was uttered. This fact can be known using very different kinds of causal links (or modes of presentation). It can be known and reported from the first-person perspective through introspection3 and introspection4 (see section 2.2), that is, via a PMIR, which has a first-order PSM as its object component. It can also be known and reported from the third-person perspective, for instance, through neuroscientific methods of investigatiori. First-person phenomenological self-reports are special in that the use of "I" takes place under the unique conditions of the uttering and the self-modeling system being identical (see the next question below).

Please note how the intentional content of self-consciousness may not supervene locally. Of course, we do ascribe much more than experiential properties to ourselves in using our conscious self-model as a basis for public statements. If you say, "Sometimes I am a little isolated," then the truth conditions for this statement are to be found in your social environment. Are you really? This intentional content is mediated through your PSM. The phenomenal content that you feel a little isolated, however, supervenes locally on brain properties. It might be a hallucination. You peed independent means of verification. What generates many problems in this domain is that beings like ourselves typically cannot make distinctions of this type on the level of subjective experience itself. For the transparent partition of our self-model, we are introspectively unable to distinguish between vehicle and content, or between self-experience and self-knowledge. However, the simple fact that you can understand these words as you read them demonstrates that, at least for human beings, the situation is more complicated.

Would the elimination of the subject use of "I' leave a gap in our understanding of ourselves?

We now have a better understanding of the constitutive conditions for what Wittgenstein called Subjektgebrauch ("subject use"): Only a true subject of experience, a system possessing a PSM and a PMIR, can refer to itself as a subject using "l." A conscious system possessing a system-model only, while not instantiating a phenomenal self-that is, a system possessing a "nemocentric" model of reality containing a fully opaque model of itself (see aboye and section 6.2.6)-could not use "I" to refer to itself as a subject of experience. The reason is straightforward and simple: such a system is no subject of experience. It has a model of reality functionally centered by a system-model, but not a consciously experienced self forming a genuine center on the level of phenomenal experience. However, we could easily adopt a terminological convention labeling this type of system as a phenomenally selfless subject of experience.

Such a system could still use the indexical expression "I." However, it would only be capable of the object use of "I," because this use would be mediated through an opaque system-model only-a globally available model of the system as an object and not through a phenomenal seIf. Such a system could truly refer to itself as "the speaker of this sentence" by using "I" leaving nothing out, because this would precisely be the way in which it would also internally represent itself: not via a phenomenal self as subject, but as a system currently generating speech output. It is easy to imagine a machine satisfying the set of constraints sketched aboye: an artificial or postbiotic system that is conscious, but has no phenomenal self, only an opaque self-model. If this system were to use "I" when communicating with us, we would therefore be justified in regarding it as an object, and not as a subject. For instance, it could never truly suffer, because it could not phenomenally own its pain states (see aboye). In this context, it must also be noted that it may make a difference if something models or refers to itself as an object only, or as a living object (cf. the discussion of Cotard's syndrome in section 7.2.2). But the deeper question in the background is whether anything that has neither a PSM nor, afortiori a PMIR could ever count as a speaker of a sentence.

Think of an unconscious patient during a prolongad epileptic automatism referrin to himself using "I" (recall the case of Dr. Z, briefly citad in chapter 6, n. 23). Is he a subject?

He certainly is not a subject of experience. Is he a speaker? He certainly does not refer to properties of himself, which were first representad as contents of his globally available (ie., conscious) self-model. It obviously is an (open) empirical, and not a philosophical matter to find out if such persons only blindly "shoot off" complex motor patterns using their physical speech apparatus, or if they actually retrieve soma unconscious form of self

representational content. Therefore, we can hardly decide ora philosophical grounds alone if such a patient might nevertheless be ara epistemic subject, that is, a subject of knowledge, even if he is not a subject of experience. I think he isn't. However, these thought experiments help us to further clarify what is the standard case: You can only linguistically refer to yourself via your conscious self-modal. The neurocomputational tool has to be in place before the linguistic tool can start to operate.

Speech is action. You peed an internal instrument that makes self-related information globally available for the flexible control of action before you can enter into external communication. This instrument is the PSM. As in the majority of cases when you are referring to information generating a phenomenally transparent form of representational content in the PSM, you do not consciously experience yourself as referring to a content when using "I" but to yourself, that is, to an object which is a subject. This, or so I propone, is the way in which attitudes and reference de se are internally modeled, else we could not understand what we are doing while using "I" Second, in all standard situations the phenomenal self is a representation of something which, although certainly possessing an objective, physical body, is in its essence a subject-a being that constantly catches itself in the act of knowing. In this way the subject use of "I" is anchored in an automatic, subpersonal process of phenomenally modeling oneself (a) as a subject, that is, as the fixed origin of a first-person perspective, and (b) transparently.

Would the elimination of the subject use of "I" leave a gap in our understanding of ourselves? There are at least two relevant readings of "understanding" at this point: individual self-understanding and theoretical self-understanding. They have to be distinguished. Consider the first case, in which an individual system possessing a PSM has stopped (or never even started) using "I" to refer to itself as subject. Let us assume it has not done so for ideological reasons, but truthfully. There are basically two classes of phenomenal systems I can conceive of as exhibiting this feature, namely, enlightened human beings or the kind of machine briefly discussed above.3 What both types of systems have in common is that they satisfy the necessary constraints for being conscious, while their self-model is fully opaque. They introspectively[3] recognize their self-models as representational structures, because earlier processing stages are continuously attentionally available to them (see section 6.4.3). The phenomenal property of selfhood is not instantiated by these systems. If they operate under a PMIR, they do not possess a consciously experienced first-person perspective, but only what I have termed a consciously experienced first-object perspective. Their variety of consciousness might well be biological consciousness (see section 3.2.II), but it would certainly not be phenomenologically subjective consciousness.

3. The Cotard's patients discussed in section 7.2.2 may constitute a third class. However, as these patients today are under heavy medication from the very beginning of their clinical stay, it is very hard to

assess their true neurophenomenological profile. In particular, their self-model is not opaque, but,-a plausible hypothesis-it is fully deprived of its emotional layer. Cotard's patients are neither enlightened nor machines, they are emotionally disembodied. See section 7.2.2.

At this point it becomes obvious that, yes, such beings would indeed have a very different individual understanding of themselves-at least as compared to normal humans like you and me. The reason is that they would live in an entirely different kind of phenomenal reality, a reality deeply counterintuitive lo most of us, a reality that seems phenomenally impossible to us because we cannot even imagine it-we are constitutionally unable to run the corresponding mental simulations in our brains (see section 2.3). If we were to encounter members of this class of systems, there would certainly be a deep gap lo be bridged in trying to understand them. If we were to become such systems ourselves, there would be an equally dramatic shift, not only in the overall structure of our conscious reality but also in our understanding of ourselves. A whole set of possible truths (e.g., about the nature of our self) would become unavailable, because there were no such truths-our understanding would only be an understanding of ourselves in a much weaker sense. The same would be true in the machine scenario. Let us now turra to the second reading of our initial question.

Would eliminating the subject use of "I"-say, in scientific and philosophical circles-leave a gap in our theoretical understanding of ourselves? Let us look at a recent example. Thomas Nagel (I986, p. 58ff.) has famously pointed out that an elimination of the particular first-person thought "I am TN" in favor of its impersonal truth conditions leaves a significant gap in our conception of the world. His general point is that all facts making such first-person, self-referential statements true can be expressed by third-person statements, but as Nagel argues, they cannot be replaced by them. We now have a much better understanding of how such third-person statements could loolc on different levels of description (see section 6.4.4 in particular). For instance, instead of saying, "l am feeling very happy today," we (or our selflessly conscious machine) could say something like, "The emotional layer of the PSM currently activated by the brain of this organism is in a close-to-optimal state." If this were a truthful autophenomenological report, we would be one of the selfless systems just discussed. But when Thomas Nagel developed the beautiful philosophical vision of the View from Nowhere he was not selfless at all. If Nagel had ever truly viewed the world from nowhere, then he would not have had any autobiographical memory referring to this episode. A fortiori he would not have been able to offer his readers a neo-Cartesian interpretation of this phenomenal episode.

As is well-known, the neo-Cartesian interpretation of the View from Nowhere faces serious analytical difficulties.[4] What is even more important is lo analyze the actual representational deep structure of the View from Nowhere, because this will also help us to

understand the Cartesian intuitions behind many bad arguments-and also what it was that Nagel has importantly discovered and what his true achievement is. We now have the conceptual tool kit to do so (briefly, for more extended discussions, see Metzinger I993 and I995c). What Nagel does is ask his readers to run a certain intended simulation (see section 2.3) in the conscious partition of their self-model. He then offers a philosophical interpretation of the resulting chain of phenomenal states. I claim that this interpretation is phenomenologically unconvincing.

4. First, the logical structure of the alleged perspectiva) fact is never clearly stated; see Lycan I987, p. 78f.; I996, p. 50; Metzinger I993, p. 233). Second, the objective self-which is more similar to Husserl's notion of a "transcendental ego" in his later philosophy than to the Wittgensteinian subject as forming the border of the world-in being used in Nagel's ubiquitous visual metaphor of the "taking" of perspectives immediately creates distal objects as its counterparts. In its conceptual interpretation this then leads to persisting act-object equivocations, to the freezing of phenomenal events into irreducible phenomenal individuals. Again, see Lycan I987, p. 79f ; I996, p. 5I f). Third, upon a closer look Nagel's concept of an "objective self" is inconsistent. It is not a mental object anymore, because the concept of mentality was introduced via the notion of a perspective as referring to subjective points of view and their modifications (see Nagel I986, p. 37). "Self," however, is a mentalistic term par excellence. Norman Malcolm has pointed out how an aperspectival objective self would be a "mindless thing" because in its striving for objectivity it would have distanced itself so radically from the point oí view of the psychological subject that it could no longer be grasped by any mental concept (see Malcohn I988, p. I58f.). The most important mistake, however, consists in using "I" as a designator and not as an indicator in the "philosophical" reading of the relevant identity statements. There are no criteria of identity offered for the individual in question. As Malcolm (I988, pp. I54, I59) puts it: "Does this make any sense? It would if there were criteria of identity for an I. [emphasis added] ... When we are uncertain about the identity of a person, sometimes we succeed in determining his identity, sometimes we make mistakes. But in regard to the identity of an I that supposedly occupies the point of view of a person, we could be neither right nor wrong. After a bout of severe amnesia Nagel might be able to identify himself as TN-but not as I. "I am TN" could announce a discovery-but not "I am I." An important source of confusion in Nagel's thinking is his assumption that the word "I" is used by each speaker, to refer to, to designate-something. But that is not how "I" is used. If it were, then "I am I" might be false, because 'r' in these two occurrences had been used to refer to different things. Nagel's statement, "I am TN," could also be false, not because the speaker was not TN, but because Nagel had mistakenly used "I" to refer to the wrong thing. If Nagel had not assumed that "I" is used, like a name, to designate something, he would not have had the notion that in each person there dwells an I or Self or Subjectwhich uses that person as its point of viewing" (Malcolm I988, p. I59f).

What Thomas Nagel terms the objective self is a conceptual reification of an ongoing representational process. This process takes place within a perspectivally structured model

of reality in the conscious mind of his readers experimenting with the View from Nowhere. Do you still remember that, when discussing mental models in chapter 3, we said that propositional representations are instructions for constructions, because they trigger internal simulations? This is precisely what happens to you when reading Nagel: Propositional input activates a chain of phenomenal mental models in your brain. In particular, you now simulate a noncentered reality within a centered model of reality. In Nagel's case, this non-centered "conception" of the world also contains all experiences and the perspective of Thomas Nagel as well:

> Essentially I have no particular point of view at all, but apprehend the world as centerless. As it happens, I ordinarily view the world from a certain vantage point, using the eyes, the person, and the daily life of TN as a kind of window. But the experiences and the perspective of TN with which I am directly presented are not the point of view of the true self, for the true self has no point of view and includes in its conception of the centerless world TN and his perspective among the contents of that world. (Nagel I986, p. 6I)

But this is false: This inner experience, the current View from Nowhere as initiated and executed by the psychological subject TN is not contained in the "centerless conception of the world." The last phenomenal event-namely, the intended shift in perspective is not contained in the centerless conception, because this would lead to an infinite regress. However, it is very obviously contained in Nagel's autobiographical self-model-else it would not be reportable. The current perspective is not a part of reality as nonperspectivally seen by the true self Nagel postulates. The threat of infinite regress is blocked by an object formation, by introducing a metaphysical entity: the objective seIf.

Here is what really happens. A conscious, self-modeling system internally simulates a noncentered reality. This simulation is opaque, and is embedded in the current PSM: at any time you know that this is only a thought experiment, and you know that you are carrying it through. Anything else would either be a manifest daydream or a full-blown mystical experience; this is certainly not the phenomenology Nagel describes. In this phenomenally simulated reality there is a model of a person, TN (or yourself), enriched by all the properties until then only known under the PSM as your own properties. This person-model forms the object component of your PMIR; it is part of a comprehensive simulational process. In this way you generate the simulation of an "inner tird-person perspective," by forming a model of yourself, which is not a self-model, but the model of yourself as if you were only given through indirect, external sources of knowledge. It is a model of a person alone in oceans of space and time, "a momentary blip on the cosmic TV-screen" (Nagel I986, p. 6I).

This process is fully reversible. In a second step you can now reintégrate the simulated person with the transparent partition of your PSM, which, of course, has been

there all along. Like a monkey or a dolphin recognizing itself in a mirror, as it were, you discover yourself in the internal mirror of your ongoing phenomenal simulation of a centerless world, by discovering a strong structural isomorphism to one of the persons contained in this world. To this representational event you can linguistically refer by exclaiming sentences of the forro "l am TN!" in their second, "philosophical" reading. But there is no homunculus that was briefly united with the transcendental ego (Nagel's objective self) and is now hurled back onto the empirical subject. This would just be a naive-realistic interpretation of a series of phenomenal representations. The cave is empty. There is no pilot. In particular, perspectivalness was never lost; constraint 6 was satisfied all the time. The View from Nowhere is a subjective state of consciousness in the sense introduced in section 3.2.II. Naive realism creeps in at the moment one forgets about the processuality (i.e., the event character and the intended nature of the respective simulations), and the phenomenal opacity (Le., the attentional availability of the representational nature characterizing the overall process).

What is interesting about Nagel's treatment is the idea of using the "vantage point" of an individual person as a kind of window. Phenomenal representations are such windows, representations under which we interact with the world and with ourselves. Some of these representational structures are opaque, but most of them are transparent: they satisfy constraint 7. A PMIR is just such a window. To represent reality (and, in higher-order mental operations, yourself) under a PMIR makes you a subject of experience. What Nagel has discovered is a fascinating architectural feature of the human mind: We are beings who can representationally distance ourselves from ourselves and make this fact globally available through conscious experience. In the terminology proposed here, what Nagel attempts to describe is that in certain special cases not only world-models but also self-models can satisfy constraint 6, the perspectivalness constraint. They do so by integrating the PMIR as a whole, modeling the intentionality relation as an internal subject-subject relation. Of course, many philosophers in the past have targeted this property, because it is a good candidate for a representational feature that distinguishes us from all other animals on this planet. Today, we can get a much clearer understanding of this feature by describing it under a naturalistic theory of mental representation, thereby preparing for a truly explanatory contact on the level of our currently best empirical theories of the mind. Using our new conceptual tools, we can describe it as the capacity to run intended, opaque emulations of our own person in our conscious self-model. We are conscious systems that can internally simulate taking an external perspective on our own person by phenomenally being the subject and object of experience at the same time. That is, we can generate a PMIR with an opaque model of ourselves as the object component. Recall how PMIRs can metaphorically be conceived of as windows. The View from Nowhere is a very specific type of PMIR, a new window through which beings like ourselves can represent the world and themselves.

Whatever is seen through this window is globally available for the formation of long-term, autobiographical memory. Its particular strength consists in the fact that it is a window being available to us at the same time.

If even the capacity to engage in the View from Nowhere is a natural property of certain representational architectures, is it, then, really true that there is nothing special about the self-ascription of phenomenal properties, be they simple bodily sensations like a tickle or a complex intemal event like the kind of self-simulation just described? Is there nothing special left? What is special in linguistic self-reference is the identity of self-modeling and self-referring system. There is only one system in the universe which can introspectively (Le., using uniquely direct causal links) access the current content of its self-model and do so under the two conditions of internality and using both tools at the same time. Maybe some future neuroscientist can indirectly read out the content of your phenomenal selfmodel to arrive at true statements about you. It is also conceivable that his predictive power in doing so might be stronger than yours, that he could actually predict your future behavior better than you could, by introspectively reading out your self-model yourself. But he would never be able to do this under the aspect of internality: your PSM supervenes locally on the internal properties of your brain, and the causal links you employ in accessing its content are uniquely direct. The neuroscientist's model of and access to yourself could never achieve this.

These are the first two defining characteristics of phenomenal self-reference that cannot be reduced to third-person reference. Third, there is a temporal reading of internality as well. For each one of us it is true that we are likely the only being in the universe that can at the same time use the neurocomputational tool (the phenomenal self-model) and the linguistic tool (the utterance of "I"). In each subject use of "I" we causally link both tools, and we do so in an extremely small time frame. "Sameness of time" here is a weak form of identity, one determined by the scope of working memory and the way the respective phenomenal system constructs its own functional window of simultaneity (see sections 2.2 and 3.2.2). Obviously, it is conceptually as well as nomologically possible for a third-person readout mechanism to operate at the same speed as the first-person process it is targeting, but the technological probability today is negligible (see Birnbacher I995). Yet what seems strictly impossible is the establishment of causal links to a conscious person's self-model that are more direct than that person's introspective capacities. Causal proximity is maximal, because introspection is itself a part of the ongoing process of selfmodeling. The underlying reason is the physical identity of the self-modeling and self-referring system: the subject of experience and the speaker are one atrd the same system. Cali this the "principie of twofold internality." This principie governing the subject use of "I" is certainly not a metaphysical mystery, but it puts each conscious subject capable of speaking a

language in a unique position. But is this an epistemologically unique position in that it ultimately presents us with an irreducible epistemological superiority or autonomy?

Is subjectivity an epistemic relation? Do phenomenal states possess truth-values? Do consciousness, the phenomenal self, and the first-person perspective supply us with a specific kind of information or knowledge, not to be gained by any other means?

Again, let us restrict the scope of our discussion to phenomenal subjectivity as discussed aboye. The PMIR as such is not an epistemic relation, but the process of consciously modeling such a relation. This process could take place in a brain in a vat. All objects (or subjects) seen through this representational window could at any point turn out to be hallucinations.[5] Therefore, any claims to knowledge-that is, to an additional epistemic or intentional content going along with the phenomenal content in question-are in need of independent justification. Phenomenal experience is how the world appears to you and as such it is nothing more than that. In particular, please recall that stronger versions of phenomenality are likely to satisfy constraint II. All the forms of consciousness, phenomenal selfhood, and subjectivity we have so far encountered were biological forms of consciousness, satisfying the adaptivity constraint. This is to say that the neuronal vehicíes subserving this content have been optimized in the course of millions of years of biological evolution on this planet. They have been optimized toward functional adequacy. Functional adequacy, however, is not the same as epistemic justification. Certain deeprooted illusions-like "believing in yourself," come what may-may certainly be biologically advantageous. It is also easy to see how the phenomenal experience of knowing something will itself be advantageous in many situations. It makes the fact that you probably possess information globally available. However, in many situations it will, of course, be functionally optimal to act as if you possess information-even if you don't. The sarne, of course, is true of the experience of knowing that you know something. To give an example, in biological and social contexts it is frequently advantageous to deceive other conspecifics, as in playing dead or pretending not to notice the presence of a certain desired object, say, a fruit or an attractive male. Deception strategies will be most reliable if they include self-deception, that is, an adequate and appropriate PSM. Due to the transparency of the self-model, the correlated phenomenal experience will be one of certainty, of knowing that you know. As many of the case studies in chapter 7 demonstrate, unnoticed and unnoticeable phenomenal misrepresentation can occur at any time. This is particularly true of higher-order or self-directed forms of representation (Neander I998). It is important to understand how such states would not be instances of self-knowledge, but could satisfy constraint II. If, in addition, my speculative hypothesis is true, that the emotional self-model also functions to

internally represent the degree of evolutionary optimality currently achieved, then it follows that certain classes of delusional states will even be emotionally attractive to beings like us.

5. Could a brain in a vat conclude that at least some kind of representational system has to exist? Can you conelude that for the very same reason your current self-model cannot be a hallucination in its entirety? Is the existential quantifier epistemically justified? That some physically realized representational system inevitably has to exist, of course, is one of the background assumptions of any naturalist theory of mental representation. At first sight, it is intriguing to see how this might lead to a naturalized version of Descartes's cogito argument. On the other hand, as a theoretical background assumption, it needs independent support. A brain in a vat-possessing no sensors, no effectors, and no social correlates to its conscious states (see section 6.3.3)-as opposed to you and me, does not have these independent sources of verification. It could therefore never even justify the assurhption that it is a representational system of some sort.

Truth-values are something predicated of sentences. Propositions are possible truths; sentences expressing such propositions possess truth-values. Our best current theories about the representational architecture underlying phenomenal experience do not assume compositionality or propositional modularity (Ramsey et al. I99I). The brain certainly is not a medium carrying out rule-based operations on syntactically specified symbol tokens. It may sometimes emulate such operations, but nevertheless the underlying laws and regularities will be the physical laws of dynamical self-organization. Therefore, it is not currently rational to assume that phenomenal states as such possess truth-values. However, as discussed extensively aboye, a specific kind of globally available self-modeling may certainly be the centrally relevant necessary condition for language acquisition. In particular, given our representationalist background assumptions, it is hard to see how the virtual organs which today we call "states of consciousness" could have propagated and preserved their own existente across certain types of nervous systems and populations of biological systems if they had not correctly and reliably extracted information from the environment in a large majority of cases. It is hard to explain large-scale reliability without assuming knowledge.

Does the incorrigibility involved in the self-ascription of psychological properties imply their infallibility?

Let us use Richard Rorty's (I970) definition of incorrigibility as implying that, given a certain subject S believes p at t, there exist no accepted procedures at this point in time which could allow us to rationally arrive at the belief that non-p. Currently, we live at such a point in time. If you refer to the content of your phenomenal self-model, there generally is no way in which a neuroscientist could demonstrate that you do so falsely. Take p lo refer

to the content of your own phenomenal self-consciousness. Your p-reports about this content cannot be corrected. However, it is important to note how the property of incorrigibility in this sense is a historical entity. Since the phenomenal content of your self-model supervenes locally and since it may be possible to discover strict, domain_ specific and law-like regularities connecting it to its minimally sufficient neural correlate, future neuroscientists may predict the content by looking at its neural vehicle. Incorrigi_ bility is a property that p-reports may lose.

In some cases p-reports have already lost this property. Recall the example of Anton's syndrome, of blindness denial, which we discussed in chapters 4 (section 4.2.3) and 7 (section 7.2.I). From all we know today about massive hemorrhages in the occipital lobes, a highly plausible inference to the best explanation in these patients leads us to the conclusion that they have no forro of phenomenal visual content which could satisfy constraint 2. We are certainly in a position today to correct the confabulations of a patient suffering from Anton's syndrome. There exist accepted procedures that allow us to arrive rationally at the belief that the patient does not have (and, sadly, never again will have) phenomenal vision-that is, that non-p. The rise of clinical neuropsychology has supplied us with many examples of situations in which human subjects actually proved to be fal. lible in terms of their phenomenal beliefs de se (for an excellent recent discussion, see Coltheart and Davies 2000). Psychiatric disorders such as Cotard's syndrome demonstrate even more dramatic possibilities, such as, for example, existence denial (see section 7.2.2). As today there are many independent reasons demonstrating the fallibility of introspec. tive autophenomenology, it is impossible to rationally draw the conclusion from incorrigibility to infallibility.

Are there any irreducible facts concerning the subjectivity of mental states which can only be grasped under a phenomenal first-person perspective or only be expressed in the f rst-person singular?

As we saw when briefly discussing Thomas Nagel's argument aboye, subjectivity is a representational phenomenon. Facts always are facts under a specific type of representation or mode of presentation. For example, you can know about the world (and yourself) under a theoretical representation, and you can know about the world (and yourself) under a phenomenal representation. In particular, you can know about the world and yourself under a phenomenal representation that satisfies constraint 6, the perspectivalnes constraint (see section 3.2.6). In this case the knowledge you gain is phenomenally subjective knowledge. It is not the represented facts that are nonobjective; the way they are portrayed by an individual conscious brain is phenomenally subjective. Given the concept of a PMIR

discussed extensively aboye, we now have a much clearer understanding ©I what this may mean in terms of the necessary representational and functional architecture involved.

A transparent self-model and a PMIR allow us to represent the world (and ourselves) in a unique manner, involving uniquely direct causal links. They also allow us to linguistically self-refer, using "I" and the current content of our PSM as a medium in a functionally privileged manner. No other system can achieve the identity of a speaking and internally self-modeling system in the same way we do. In this way our consciously experienced and linguistically or cognitively extended first-person perspective is truly an individual first-person perspective. Our phenomenal model of reality is an individual picture. Yet all the functional and representational facts constituting this unusual situation can be described objectively, and are open to scientific inquiry. Consciousness is epistemically irreducible, but this irreducibility is now demystified, because we have a better understanding of how epistemic subjectivity is rooted in phenomenal subjectivity. In order to have subjective knowledge, you need to successfully represent reality (and yourself) under a conscious world-model that satisfies constraint 6.

Can the thesis that the scientific worldview must in principle remain incomplete be derived from the subjectivity of the mental? Can subjectivity, in its full content, be naturalized?

First, let us restrict the scope to phenomenal subjectivity again. As we have seen in the course of this book, no principled obstacles to an exhaustive representationalist analysis of consciousness, selfhood, and perspectivalness exist. This is not to say that we may not discover such obstacles in the future, or eventually do without the representationalist level of description altogether. For now, it is tenable to say that all phenomenal facts are representational facts and that phenomenal subjectivity is a representational phenomenon in its entirety. The consciously experienced first-person perspective is simply one of an infinitely large number of possibilities in which a representational system can portray reality.

Second, of course, it is not clear what subjectivity in its full content actually is. I see two major extensions of maximal relevance which have only been touched upon in this book in passing: intentional subjectivity and intersubjectivity. Intentional subjectivity is something that will not supervene locally on brain properties. It is the question of self-knowledge versus the question of self-experience. Our unconscious self-model incorporates huge amounts of information about our physical body and its relationship to the environment. In part, this information was acquired by millions of generations of our biological ancestors. As a forro of unconscious, nonconceptual, structurally embodied self-knowledge, that is, as an intentional content, it satisfies the adaptivity constraint. Then there is a wide variety of occurrent self-representational states, for instance, when thinking about

ourselves. This self-directed intentional or representational content is something different from the phenomenal character by which it may or may not be accompanied.

Arguably, one can even have a weak, passive version of unconscious thoughts about oneself, for instance, during phases of non-REM (NREM) sleep. Dreaming and REM sleep are incompletely correlated, because up to 30% of REM awakenings do not elicit dream reports, whereas up to I0% of NREM awakenings do lead to reports about complex forros of mentation, which, interestingly, have a more cognitive type of content than the usual phenomenal dreams that many of us recall in the morning (for details and references, see Nielsen 2000; see also Solms 2000). Therefore, any more general and comprehensive theory of subjectivity will have to do justice to such nonphenomenal types of mental content as well. They are outside the scope of this work, but they will be an important part of a fuller understanding of mind and subjectivity.

Second, the history of consciousness has not stopped with the PMIR. It has already taken the step from the phenomenal first-person perspective to all other forms of conscious experience from which our rich, social reality eventually emerges: the phenomenal experience of the "you," the "he/she/it," the phenomenal awareness of "us," "you," and "them." I have in this book only focused on the PSM and PMIR, because I see them as the decisive link between personal and subpersonal truths about the human mind, and because I think that they actually may Nave been the crucial step from biological to cultural evolution. But subjectivity in its full content will certainly have to include not only subjectobject relations, but subject-subject relations as well. It will also have to include subject-group relations. To give just two examples: There will be a phenomenal (and intentional) representation not only of "mineness" and ownership. There will also be "usness" and "themness"; there will be a mental representation of the first- and third-person plural. The notion of a consciously experienced perspective is greatly expanded if we want to do justice to such important facts. The phenomenal first-person perspective now looks like the functional or representational foundation of the conscious first- and third-person plural perspectives. Obviously, we are today very far from being able to furnish anything in terms of an empirically anchored, rigorous, and conceptually convincing analysis of the kinds of mental representations involved in all of these target phenomena. It is therefore better to be modest, start at the very beginning, and try to understand phenomenal subjectivity first.

Do anything like "first-person data" exist? Can introspective reports compete with statements originating from scientific theories of the mind?

The popular notion of "first-person data" is a metaphor, just like the notion of a "first-person perspective." In both cases, an ill-defined but intuitively attractive meaning results from the combination of two precursor concepts originating in very different domains. In

the latter case, we fuse a semantic element of grammar theory (ie., associated with linguistic ascriptions of certain properties in the first-person singular) with a semantic element related to the phenomenology of visual experience, in particular to its geometry (as a contingent matter of fact our visual model of the world is centered around a point of view: distant objects appear smaller than those in our vicinity, parallel fines converge at the horizon, etc.). In the first case, we fuse the same semantic element with a concept borrowed from the theory of science. In doing so we put the notion of "data" to an extended usage, which unfortunately runs the great risk of simply being empty. First, data are things that are extracted from the physical world by technical measuring devices like telescopes, electrodes, or functional MRI scanners. There is a well-defined and public procedure, which certainly has its limitations, but which can be and is being continuously improved. Second, data generation inevitably takes place among groups of human beings, that is, within scientific communities open to criticism and constantly seeking independent means of verification. Data generation is, by necessity, an intersubjective process. First-person access to the phenomenal content of one's own mental states does not fulfill these defining criteria for the concept of "data." My politically incorrect conclusion therefore is that first-person data do not exist.

Of course, maximizing phenomenological plausibility is of the highest priority for any theory of consciousness, the phenomenal self, and the first-person perspective. In this book I have tried to develop an alternative strategy, namely, by maximizing the degree of phenomenological constraint satisfaction. As the reader may remember, phenomenological constraints were always the first constraints from which I started (for the only exceptions, see sections 3.2.II and 6.2.8). The advantage of this somewhat weaker procedure is that you get all the heuristic power from first-person descriptions without being driven to naive-realistic assumptions and the stipulation of mysterious, nonpublic objects. In particular, you can define networks of constraints that can be continuously refined on lower levels of description, while at the same time allowing you to search for domain-specific, consistent solutions. You can take phenomenology seriously without running into all of its traditional problems.

The epistemological problem regarding phenomenological, first-person approaches of "data generation" is that if inconsistencies in two individual "data sets" should appear, there is no way to settle the conflict. In particular, the phenomenological method cannot provide a method of generating any further growth of knowledge in such situations. Progress ends. This is a third defining characteristic of the scientific way of approaching reality: there are procedures to settle conflicts resulting from conflicting hypotheses. Epistemic progress continues. The same is not true in cases where two experiential subjects arrive at conflicting statements like "This is the purest blue anyone can perceive!" versus "No, it isn't, it has a faint but perceptible trace of green in it!" or, "This conscious experience of jealousy shows

me how much I love my husband!" versus "No, this emotional state is not love at all, it is a neurotic, bourgeois fear of loss!" The advantage of the constraint satisfaction approach is that we can turn such discoveries into new and differentiated constraints themselves. Any good theory of consciousness now has to explain how such truthful but conflicting autophenomenological reports are possible, and in which cases they will emerge by necessity. Inconsistencies in reports lead to progress by differentiating the constraint landscape.

Can introspective reports compete with statements originating from scientific theories of the mind? Yes, they can, and they should. But please note how any such competition is relative to our interests: What do these statements compete for? If our agreed-on goal is predictive power, then it certainly is possible to assign such a power to first-person autophenomenological statements like, "l will always be able to discriminate my purest blue from your purest blue!" or, "You will never be able to consciously experience a colored patch that exhibits red and green presentational content at the same time, while fully satisfying constraint I0, the homogeneity constraint!" Such statements are always statements about publicly observable future behavior. They make predictions. In the first case, the experiential subject may be the winner; in the second case, science may make the better predictions (Crane and Piantanida I983). Right now, first-person predictions of one's own future behavior-which are invariably based on introspectively accessing the content of one's PSM-are much better and more reliable than third-person predictions. This situation may change, as we learn more about potential divergences or dissociations between the phenomenal and the functional, behavior-driving layers in the human selfmodel or about the evolutionary advantages of self-deception. What will not change is the remaining, and deeper, philosophical question. It is the issue of what actually it is that introspective reports and statements originating from scientific theories of the mind can compete for.

The true focus of the current proposal, however, was phenomenal content, the way certain representational states feel from the first-person perspective. Does it help to shed new light on the historical roots of certain philosophical intuitions like, for instante, the Cartesian intuition that I could always have been someone else; or that my own consciousness necessarily forms a single, unified whole; or that phenomenal experience actually brings us in direct and immediate contact with ourselves and the world around us? Philosophical problems can frequently be solved by conceptual analysis or by transform ing them into more differentiated versions. Sometimes these new versions can be handed over to the sciences. Arguably, the problem of consciousness can be naturalized by transforming it into an empirically tractable version. However, an additional, complementary, and equally interesting strategy consists in attempting to also uncover their introspective roots. A careful inspection of these roots may help us to understand the intuitive force behind many bad arguments, a force that typically survives their rebuttal. I will therefore

supplement my discussion by having a closer look at the genetic conditions for certain introspective certainties. But let us look at experiential content first.

What is the "phenomenal content" of mental states, as opposed to their representational or "intentional content?" Are there examples of mentality exhibiting one without the other? Do double dissociations exist?

The representational or intentional content of a mental state is what this state is directed at. The important point is that this is a relational and an abstract property, not an intrinsic property of the physical state carrying the content, and that this is true of self-representational intentional (or, in traditional parlance, "reflexive") content too. A self-model gains its intentional content by being directed at the system as a whole, at the system within which it is activated. It has a single intentional object. The same is true of what I have called the PMIR, the phenomenal model of the intentionality relation itself. If we exclusively view it as a representational structure, then it is directed at certain classes of subject-object relations. It is directed at the fact that the system is currently attending to a certain visual object, or thinking about something specific, or an agent pursuing a certain goal state, or in communication trying to understand the thoughts of another human being. The second important point is that everything that has an intentional content can misrepresent either the external world or the representational system itself. For instance, it could misrepresent the fact that it actually is pursuing a certain goal state.

Then there is phenomenal content. It is a special form of intentional content, namely, in satisfying the constraints developed in chapters 3 and 6. Importantly, there are now many different degrees of phenomenality: an intentional content can be more or less conscious. Our new conceptual tool kit allows us to describe very many levels of subjective experience, thereby doing justice to different domains and a large number of phenomenological constraints. Phenomenal states and events are a proper subset of intentional states and events. For a given minded being, and a given point in time, this subset may be empty. To give another illustrative example, think about the case of NREM sleep mentation briefly mentioned aboye. Unconscious thinking taking place in the NREM phase of nocturnal sleep certainly is representational activity, it may be misrepresentational activity, but it is not globally available for action control, for attention, or for selective metacognition (as a matter of fact, this is demonstrated by perserverative characteristics). Raffman qualia and Metzinger qualia (see section 2.4.4) are further examples of intentional content, which is only weakly conscious, because it only satisfies the globality constraint for attention, but not for cognition, and in the latter case not even for action control.

These forms of presentational content are intentional, because-although no simple or systematic one-to-one mapping with any kind of physical property is possible-they are, in

the sense of an ancient, approximative, and unreliable teleological function, directed at certain properties, of certain objects, in the ecological niche of certain animals. As we saw in section 3.2.II, these properties do not have to be surface properties, they can be hidden physical properties like the fact that certain types of young leaves are richer in protein (Dominy and Lucas 200I). How they appear to us, millions of years after they acquired this first function in our distant ancestors, is another matter. Please note that the same may be true of self-modeling. Much of it may actually be unconscious or weakly conscious. Much of it was acquired millions of years ago by our distant ancestors. And a large parí of the human self-model may originally have been directed at target properties, which were properties of our ancestors, but are not properties of us anymore. Self-perception may frequently not correspond to the internal stimulus itself, but to an ancient internal context, to the probability distribution of its possible sources millions of years ago. In particular, it is an empirically plausible assumption that the largest portion of the selfmodeling that causally drives our behavior actually takes place unconsciously. This insight reaches across a large range of cases, from the internal emulation of fast, goal-directed movement (see, e.g., section 7.2.3) to social cognition (see section 6.3.3). The phenomenal self-model is only that partition of the mental self-model which satisfies a certain subset of items in our flexible catalogue of constraints. For the PMIR, the situation is more difficult: Is it plausible to assume that unconscious modeling of intentional directedness itself takes place as well? Is there a non-phenomenal MIR, or is perspectivalness truly the hallmark of conscious experience and conscious experience only? Interestingly, we now have a refined version of Brentano's ([I874] I973) original point. Fortunately, as Brentano would be delighted to hear, this is a question that can now be settled by empirical research and not by philosophical speculation.

Double dissociations do not exist. There certainly is unconscious intentional content. A lot of it. But in ecologically valid standard situations there is no conscious state that is not a representational state in some way (for a nonstandard situation, cf. the abstract geometrical hallucinations discussed in chapter 4; see also figure 4.I for one beautiful example of purely phenomenal content). As long as we choose to operate on the representational level of analysis at all-and this may change-there is no example of phenomenal content that is not also directed at some target object, property, or relation. Please note that this does not mean that the experiential subject has to have the slightest clue about what the intentional object of his or her experiences actually is. In many cases, for example, in living through diffuse feelings and emotions (like jealousy), the original intentional object may be millions of years away. It may not exist anymore. The original representandum may be something that was only present in the world of our distant ancestors. In particular, as we have learned from our discussion of the transparency constraint, Mother Nature has until very recently not cared to let any of us know or

experience the fact that there is something like intentionality at all. Only recently did we become able to discover and represent the fact that we actually have minds. The PMIR may actually be our first attempt at internally making this very fact accessible to ourselves.

Please also note how a phenomenal state may be only weakly representational. Under certain theories of mental representation, for example, those that describe degrees of statistical dependency (see Eliasmith 2000) or covariance, one and the same type of phenomenal state may satisfy certain constraints for intentionality (e.g., accuracy) to differing degrees on different occasions. It may be more or less representational. But even if it is pure appearance, a misrepresentation in its entirety, it still has an intentional object: it is directed at something. The interesting question is whether in such cases its directedness consists in more than the fact that it is integrated into a PMIR. However, as it has not been my goal to offer a general theory of mental representation, this issue is clearly outside the scope of the current investigation.

How do Cartesian intuitions, like the contingency intuition, the indivisibility intuition, and the intuition of immediate givenness, emerge?

The degree of intuitive plausibility for a given theory results from the degree of phenomenal possibility associated with it (see section 2.3). Theories describe worlds. Conscious experience models worlds. Beings like ourselves experience all those theories as intuitively plausible that describe worlds that can be phenomenally simulated by us. These worlds then strike us as possible phenomenal experiences we might have-because we can internally model them. Therefore, this concept of possibility is always relative to a certain class of concrete representational systems, each of which possesses a specific functional profile and a particular representational architecture. Human beings may-and do-differ in what they can imagine, in what classes of worlds they can consciously simulate, and in what they find intuitively plausible. We cannot imagine the thirteen-dimensional shadow of a fourteen-dimensional cube or the continuum of space-time, because the visual cortex of our ancestors was never confronted with this type of object and because the brain's global model of reality based on three spatial dimensions and one distinct, unidirectional temporal dimension sufficed for surviving in what was our biological environment.

Moreover, intuitions change over a lifetime: Even within one individual the internal landscape characterizing the space of possibilities may undergo considerable change. It is also important to note how the mechanisms of generating and evaluating representational coherence employed by such systems have been optimized with regard to their biological or social functionality, and do not have to be subject to the classic criteria of adequacy, rationality, or epistemic justification in the narrow sense of philosophical epistemology. Briefly, in our own case, the set of phenomenally possible worlds is related directly to the

set of nomologically possible worlds, but only indirectly to the sets of logically and metaphysically possible worlds.

The contingency intuition is the intuition of thinking " I could always have been somebody completely different!" Proponents of essentialist theories of subjectivity traditionally have been guided by this intuition, by the beautiful-and certainly emotionally attractive-idea that there must be something about myself which has nothing to do with any of my ever-changing, objective, and observable properties. As a subject I just cannot be identical with my physical body or any of its more complex and abstract properties-at least this identity is not a necessary identity. It could be broken. I will not go into the long history of this philosophical intuition here, as I have already presented one recent example in this chapter. If Thomas Nagel (I986, p. 6I) says "Essentially [emphasis added] I have no particular point of view at all ... ," we have exactly this situation: There is an essence, the objective self, which is only contingently united with the history of a certain physical person by the name of T. N. and its individual perspective. However, as we saw aboye, the phenomenology of the View from Nowhere is one in which the original PMIR is never really lost, and its conceptual interpretation is flawed. What Nagel describes is not a mystical experience-the Great View from Nowhere-but just an ordinary thought experiment.

The undisputed fact is that all of us can imagine having had a completely different set of public and phenomenological properties, say, those of Immanuel Kant. However, in doing this we only open a certain new partition of our phenomenal space and activate a fictitious self-simulatum, more or less completely portraying us as possessing those properties of Immanuel Kant currently known to us. What we construct is a cognitive firstperson perspective (see section 6.5.2), a PMIR with the simulated person as its object component. This is a first-person state, a conscious model of reality satisfying the perspectivalness constraint, constraint 6. In order to really consciously simulate a world in which we would have been Immanuel Kant, the Kant-model would have to be the transparent subject component of this PMIR. As most of my colleagues know, in philosophical departments around the world, such systems actually do appear from time to time-systems in which a constant attempt at counterfactual self-simulation has gotten out of control, and which, due to a now highly afunctional self-model, actually believe they are Immanuel Kant, and even experience themselves accordingly. What these sad cases of delusion show exactly is that, in these cases, there is no overarching essence, no subjective core state or genuine phenomenal identity anymore.

But isn't it true that we can actually imagine much more than only being Immanuel Kant, namely, being Immanuel Kant, including his PMIR? Isn't the View from Nowhera more like social cognition, in that its object component truly is another subject, including the representation of a second PMIR (see section 6.3.3)? Yes, certainly. But social cogni, tion is a first-person process, even if directed at a second, fictitious self. Being someone is a

phenomenal property determined by the locus of attentional, volitional, and cognitive agency, as represented under a transparent self-model (cf. our discussion of OBES in section 7.2.3). If you represent your alter ego as being a subject, as having a PMIR, thCI this subject component remains opaque. In nonmystical and nondeluded states, the ftrst, the original person-model invariably is the transparent one: You know that you do the imagining-at least this fact is globally available to you at any time. So what you imagine is never being Immanuel Kant, even if you phenomenally simulate him as subject, possessing a first-person perspective of his own. You cannot simulate him as self. So the surprising answer is that the contingency intuition, at closer inspection, is not even based on a phenomenal possibility. As a philosophical claim, it is based on bad phenomenology. And this seems to contradict my introductory claim that intuitive plausibility goes along with phenomenal possibility.

There must be a second factor, which, for beings like ourselves, makes it attractive to believe in essentialist interpretations of the sort Nagel offers. Death denial may be this factor. In chapter 2 we saw how the process of phenomenal simulation needs a heuristic which compresses the vastness of logical space to two essential classes of "intended realities," that is, those world models causally conducive and relevant to the selection process. The first class is constituted by all desirable worlds, that is, all those worlds in which the system is enjoying optimal external conditions, many descendants, and a high social status. A world in which we are potentially independent of our physical bodies, a world in which individual survival is, in principle, possible is certainly a desirable world for beings like us. Individual survival is one of the highest biological imperatives burned into our emotional self-model: This is where we come from. Even if we cannot really carry out the phenomenal simulations needed, what we may feel is that such simulations would satisfy constraint II, the adaptivity constraint. They are emotionally attractive, and that is why beings like us are all too ready to jump to certain inaccurate descriptions, and, in particular, to related assumptions about metaphysical possibility. False beliefs certainly can be adaptive. If mental health is defined as the integrity and stability of the self-model, then some types of false beliefs may even be conducive to mental health. So the answer is that there is a second functional factor: intuitive plausibility not only goes along with phenomenal possibility per se, but with adaptivity as well. But what, in this special case of Cartesian and essentialist intuitions about the phenomenal self, is the adaptation directed to? Here is a speculative hypothesis: it may be directed to a recent change in our selfmodel. This change may have been the split into a transparent and an opaque section, and, in particular, to the entirely new situation that this split made the fact of our own mortality cognitively available. We paid a high price for becoming cognitive subjects, and essentialist fantasies may be an attempt to minimize this price as much as possible. More about this in the final section.

In the Sixth Meditation, Descartes attempted to make it appear as an immediately evident truth that "there is a vast difference between mind and body, in respect that body, from its nature, is always divisible, and that mind is entirely indivisible." This is a different situation. Here, we clearly see how intuitive plausibility is rooted in a phenomenal necessity. It is not possible that our minds are not "absolutely one and•entire" (as Descartes put it), because beings like ourselves cannot run the corresponding phenomenal self-simulations. As noted aboye, one of the most fundamental functional constraints on the PSM is that the system currently operating under it cannot intentionally split it or dissolve it. Have you ever tried? And in confusing phenomenal with logical modalities, it may therefore appear that what is not not possible is necessary, in some sense of a priori necessity which now has to be explained. The existence of a single, unified self may even appear as metaphysically necessary. But it isn't. As the neurophenomenological case studies presented in the preceding chapter demonstrate, there are many kinds of conscious but highly fragmented self-modeling of which it is true that they are strictly impossible to imagine. Can you imagine what it is like for a Cotard patient to be absolutely certain that she does not exist? Can you imagine how it is for a schizophrenic to have alien thoughts penetrating his mind? For healthy people there simply is no way to imagine this, clearly and distinctly. And there may again be a deeper, teleofunctionalist reason for this obvious fact.

Put very simply, we are not supposed to imagine pathological situations, because dissociatíve self-simulations endanger the functional integrity of the organism as a whole. After all, the self-model, in its deeper functional core, is an instrument used in homeostatic self-regulation, a tool to make the individual process of life as coherent and stable as possible. Too much playing around in its conscious offline section could eventually put the elementary processes of elementary bioregulation at risk. It could make you sick. Cotard's syndrome, schizophrenia, and other identity disorders are highly maladaptive situations-they must be avoided at all costs. All situations in which the conscious selfmodel, in one way or another, portrays the system as falling apart into two or more parts (again, recall the case study on OBEs in section 7.2.3) are usually situations in which the organism simply is in great danger of dying. The integrity of the organism as a whole is at risk. Therefore, the corresponding self-simulations are emotionally unattractive; they can even cause a fear reaction. A related fact is that many people find it distressing, or painful, or threatening to seriously try to understand patients with severe psychiatric disorders or to be among them for a long period of time. The Cartesian intuition of the indivisible self, the Kantian notion of the transcendental subject-the reassuring idea of the "I think" that can, at least in principle, accompany all my conscious Vorstellungen-are rooted in this feature of our representational architecture, in the functional inability of the system as a whole to split its PSM. This is why monist or naturalist theories of subjectivity inevitably strike us as deeply

counterintuitive, and frequently even as emotionallY unattractive. Of course, all this is no argument showing that Descartes and Kant may not have been right.

What about the Cartesian picture of a sensory-cognitive continuum, the idea that we can directly gain knowledge about the world and about ourselves through the process of conscious experience? Is there epistemic immediacy going along with some of the contents constituting consciousness, the phenomenal self, and the first-person perspective? Here, our answer can be brief. For all conscious mental content satisfying the transparency constraint in the way introduced here (see sections 3.2.7 and 6.2.6), it is an obvious and necessary characteristic that this content will be experienced as immediately given. This characteristic is a phenomenological feature in its entirety. Phenomenal immediacy does not entail epistemic immediacy. Every form of phenomenal content is in need of independent epistemic justification, and this, of course, is also true of the conscious experience of apparently direct perception, direct reference, direct knowledge, and so on. Our neurophenomenological case studies in chapters 4 and 7 provided examples of a wide range of states which are, in a way that is unnoticeable to the subject of experience, epistemically empty constructs and at the same time characterized by the phenomenology of certainty and direct, immediate knowledge. In particular, it is today empirically implausible to assume that mental contents satisfying at least constraints 2, 3, and 6 could not be based on complex, physically realized, and therefore fallible and time-consuming processes of information processing. There simply is no such thing as epistemically immediate contact to reality. What there is is an efficient, cost-effective, and evolutionary advantageous way of phenomenally modeling reliable representational contents as immediately given.

In chapter I, I pointed out how the human variety of conscious subjectivity is unique on this planet in being deeply culturally embedded (see also Metzinger 2000b, p. 6ff.), namely, through language and social interactions. It is therefore interesting to ask how the actual contents of experience change through this constant integration into other repre sentational media, and how specific contents may genetically depend on social factors.

Which new phenomenal properties emerge through cognitive and linguistic forms of self-reference? In humans, are there necessary social correlates for certain kinds of phenomenal content?

As we saw in section 6.4.4, cognitive self-reference is not something taking place in a linguistic medium, but a very special way of higher-order self-modeling. This new way of self-modeling, in particular when internally emulating logical operations involving rule-based transformations over discrete symbol tokens, and so on, has been a major break-through in biological intelligence. It has made abstract information globally available, for example, information about what is logically coherent or information about the fact that

there is a difference-and often some kind of relation-between reality and representation, and it arguably has even enabled us to form the concept of truth. After language was available, we could not only communicate about all these new facts and concepts but also proceed to publicly self-ascribe them to us. And this, of course, brought about new phenomenal properties, because it dramatically changed the content of the PSM.

We started to consciously experience ourselves as thinkers of thoughts and as speakers of sentences. We started to think about ourselves as thinkers of thoughts and as speakers of sentences. And we again consciously experienced this fact, because it brought about a change in our PSM. We started to talk to each other about the surprising fact that we-y very likely as opposed to most other creatures we knew-are thinkers of thoughts and speakers of sentences, and that we know about this fact, because we consciously experience it. Or do we? We mutually started to ascribe the property of being experiencing, thinking, communicating beings to ourselves and to each other, and because the difference between reality and representation was already available to us, due to the phenomenal opacity of our cognitive self-model, we were aware that such ascriptions might actually be false. We started to discuss matters. We disagreed! Probably, philosophy was bom at this stage. I will not speculate here concerning the more fine-grained representational architecture that must have been involved in these first steps. I just want to point out how this chain of events-self-modeling systems now starting to mirror each other not only through their motor systems (see section 6.4.4) but also through the opaque sections of their minds and through external use of symbols-has brought about two fundamental and highly interesting shifts in the phenomenal content of the human-self-model.

First, we could begin to experience ourselves as rational individuals. Because it was now possible to consciously model the process of forming coherent thoughts according to some set of abstract, logical rules, we could also form the concept of a being that actively strives for this kind of coherence. We already knew that we had emotions, that we were biological beings with needs and urges, following the logic of survival. Now we started to see that for us there was another-possibly a conflicting-way of following a certain logic. This time it was the logic of intellectual integrity, of making the opaque partition of your PSM as coherent as possible. It was the logic of having as few conflicting thoughts as possible. It was the logic of preserving truth. And it also was the logic of rational agency, of pursuing one's own goals by ordering them into a consistent hierarchy, and the logic of trying to gain knowledge in order to achieve these goals, to continuously minimize the difference between mental representation and reality already discovered. We experienced ourselves as individual beings that, at least to certain degree, also were rational subjecti. Although our PSM was now continuously split into an opaque and a transparent partitiott, our brains somehow managed in that it remained one single representational structure and

thereby allowed us to phenomenally own this new property of ourselves. Later, we even found a new linguistic concept to describe this new property. We called it personhood.

Second, we could now also begin to make the fact that we are social subjects globally available for attention, cognitive processing, and action control. In particular, we coul start to consciously experience the fact that, as social beings, we were not only related tO each other through the common logic of survival, that is, through our bodies and our emotions, but as rational persons. The new phenomenal property of personhood could now start to unfold its functional profile. Because it made this radically new information globally available for deliberate action control, for linguistic report and communication, for self-ascription and critical discussion, it was now possible for us to also share this information. Our PSM allowed us to pool our cognitive resources. Of course, as we have seen in this book, intersubjectivity starts on the unconscious level and is later mediated through entirely nonconceptual and nonpropositional levels of the PSM. But now we had acquired a PSM of ourselves as persons, as rational individuals. Rational individuals are capable of rational intersubjectivity, because they can mirror each other in an entirely different manner. At this point a whole cascade of functional transformations, of instantiations of new functional properties through the global availability of new representational contents, unfolded explosively.

The new kind of PSM enabled the construction of a new type of PMIR. Its object component could now be formed by other subjects, this time phenomenally modeled as individual thinkers of thoughts and as rational, self-conscious bearers of strong first-person phenomena. Person-te-person relations in the true sense of the word could be consciously modeled. I will not go into any further details at this point, because I think it is already clear how the human PSM was the decisive neurocomputational tool in the shift from biological to cultural evolution. Let me just narre what I think is the essential high-level property, the central functional feature, in which this transition eventually culminated. Through our extended PSMs we were able to simultaneously establish correlated cognitive PMIRs of the type just described. Two or more human beings could now at the same time actívate cognitive PMIRs mutually pointing to each other under a representation as rational subjects. And the correlated nature of these two mental events, their mutuality and interdependence, could itself be represented on the level of global availability. We were able to mutually acknowledge each other as persons, and to consciously experience this very fact.

The concept of a "person," however, does not simply refer to some complex, but objective representational property. Personhood cannot be naturalized in a simple and straightforward way, because the concept of a person contains domain-specific and semantically vague normative elements. Why is this so? Persons never are something we find out there, as parts of an objective order. Persons are constituted in societies. If conscious selfmodeling systems acknowledge each other as persons, then they are persons.

But, as I pointed out in chapter I, conscious experience is a culturally embedded phenomenon (see also Metzinger 2000b). This is true of complex phenomenal properties like personhood too. From the perspective of the humanities, it is therefore centrally important to gain a more precise understanding of the neurocognitive and evolutionary, of the functional and representational conditions of possibility governing the appearance of personhood and successful intersubjectivity. And this is what I have tried to prepare for in this book. But how the emergence of phenomenal personhood and the mutual processes of "mirroring" each others' conscious personhood described aboye are then interpreted in a given social context is quite another matter. Answers to the question of how they should be interpretad may vary from society to society, from subculture to subculture, or from one historical epoch to another. And even given our own context, they may certainly change as we now learn more about what in our brains and in their biological history actually brings them about.

The question about the necessary social correlates of consciousness, selfhood, and high level perspectivalness today has become a predominantly empirical issue. However, at the risk of being tedious, one has to be clear about the underlying metaphysical principle of local supervenience. To bring out this point, let us take an example not yet involving the complex neurophenomenology of rational intersubjectivity. Let us choose something much more simple and beautiful: the conscious experience of very briefly catching a glimmer in the eye of another human being; a glimmer that, in the fraction of a second, lets you discover not only that, this person likes you but also how much she likes you, and the conscious experience of realizing, at what appears to be the same time, that you yourself must have been giving the same signal a moment ago. SMT makes the claim that even for consciously experienced intersubjectivity of this type, it is true that an appropriately stimulated brain in a vat could activate the same phenomenal content. This claim may strike you as bizarre. But it is easy to understand what this claim amounts to, and what it does not amount to as well.

First, in standard situations, phenomenal content is a special kind of intentional content. In order to understand how it could ever come about, what its role in the psychological ecology of its bearers actually is, how this role has been shaped by the history of their biological ancestors, and so forth, you have to give a much more extensive representationalist, teleofunctionalist, and eventually even neuroscientific analysis. You peed all the subpersonal levels of description on which I have operated in chapters 3 and 6 in order to arrive at a truly informative analysis of consciously experienced social cognition. Local supervenience is just a (rather weak) metaphysical claim, one that in various ways assumes asymmetrical bottom-up dependency without reducibility. One of the weaknesses of supetvenience is that it is not an explanatory relation. In saying that phenomenal intersubicctivity supervenes locally on individual brain properties, you are not

saying that social and intersubjective knowledge is determined by internal and contemporaneous brain properties as well. You are not even contributing to a deeper understanding of why this knowtedge had to be mediated by conscious experience.

At any given point in time, phenomenal content supervenes locally on properties individual brains. "Points in time" are physical entities, individuated from a thid-person perspective. Brain properties are fully embedded in the causal network of the physiced world, and information processing in the brain is a time-consuming process. Therefore, social cognition and the conscious processing going along with it are time-consuming processes as well. There is no such thing as temporal or epistemic immediacy on any subpersonal level. However, there may certainly be phenomenal immediacy, for instance, in the situation described aboye. In the way the brain individuates points in time (see section 3.2.2) it is certainly conceivable that the spark of sympathy shooting back and forth between two conscious human beings may be experienced as an instantaneous spark. It may be transparently represented as one single event, taking place in one single moment, but bridging the gulf between two individuals. Phenomenologically, lightning strikes and mutually unites two phenomenal selves-this is the "affective dissolution of the self" mentioned earlier. Because it involves loss of control over and transient dissolution of the emotional self-model, the experience of catching each other in the act of falling in love is a little bit like dying, and also a little bit like going insane. As a phenomenal content, this event supervenes locally. As a nonconceptual form of intentional content it doesn't.

At least in some cases, becoming friends or falling in love is a process of knowledge acquisition. We cannot even begin to adequately understand it if we do not understand the information that it makes globally available for the system as a whole, plus the representational and functional role such an event plays for the now coupled system of our two self-modeling systems. And on this level of analysis it is all too obvious that many forms of conscious experience-which, after all, are a special form of intelligence too-possess necessary social correlates (e.g., see section 6.3.3). What these correlates correlate with is invariably the self-model of the other organism. The self-model is what built the functional bridge from individual cognition to social cognition, from first-person intelligence to the pooling of resources in a species. Such resources can be intellectual, but they can also be emotional or motivational. The first step in understanding social cognition therefore consists in developing an acquisition history for this new virtual organ, in telling a comprehensive developmental and evolutionary story about the PSM and the PMIR in particular. Metaphysically, individual occurrences of phenomenally experienced intersubjectivity supervene locally. But-and this is the answer to our original question-if we want to understand how they can satisfy the adaptivity constraint, we will need a greatly expanded explanatory base.

In chapter I I also promised answers to a set of questions concerning the relations between certain phenomenal state classes or global phenomenal properties. Let us now, briefly, look at the answers.

What is the most simple form of phenomenal content? Is there anything like "qualia" in the classic sense of the word?

Simplicity is representational atomicity. What appears as simple and as strictly indivisible is always relative to the representational architecture actually generating this content (see section 2.4; see also Jakab 2000). There are two readings of this answer: First, simplicity is relative to the internal readout mechanism employed by the brain in activating a globally available presentatum of, say, magentaII and integrating it into short-terco memory, then making it an element of the object component of the PMIR; second, simplicity is relative to the theory describing this mechanism and to the constraints which this specific theoretical solution attempts to satisfy. Let us remain with the first case. It makes a difference if presentational content is made available for introspectionI or for introspection2; that is, if it becomes conscious by being available for attentional processing and behavioral control only, or by being available for the formation of enduring, concept-like mental structures as well. Many conscious color experiences, for example, are simple phenomenal states in tercos of being attentional atoms, but not cognitive atoms. Therefore, they are ineffable (Raffman I995) and simpler than qualia in the classic sense of the term.

Lewis qualia in the sense of the classic terminology (see section 2.4.I)-as a first-order phenomenal property, that is, as a recognizable, maximally simple, and fully determinate form of sensory content-arguably do not exist. For most maximally determinate sensory values it is a truism that we cannot, as Lewis originally demanded, reliably recognize them from one instance to the next. We cannot form concepts of them because, due to limitations of perceptual memory, we possess no transtemporal identity criteria. That is, they; are not cognitively, but only attentionally available from the first-person perspective. They can only contribute to the object component of a certain, restricted subset of PMIRs. The interesting fact (which some analytical philosophers all too much would like to ignore) is that not only are we unable to talk about them in a conceptually precise manner but we', cannot even think about them. Certainly something like strong Lewis qualia do exist, for example, in terms of the conscious experience of the pure colors green, blue, red, andd yellow. But it is not the most simple form of phenomenal color content. As extensively explained in section 2.4.4, Lewis qualia are much stronger than Raffman qualia, because, descriptively, they are located on the three-constraint level, and not on the two-constraint level. In this second sense, unitary hues, due to their "purity" and the resulting cognitive availability, are a richer and more complex form of phenomenal content. The more general

image that now emerges (particularly for the level of allegedly "simple" sensory processing) portrays conscious experience as a highly graded phenomenon. Phenomenal colar vision, for example, is not an all-or-nothing affair but recedes gradually into the uncon~ scious processing of wavelength information through a fine-grained hierarchy constituted by levels of constraint satisfaction. And this is a second sense in which simplicity is relative to the representational scheme under which it appears: simplicity is theory-relatiti'e. Take the current investigation as an example.

In chapter 2 I used an excessively simplistic and crude example of constraints that can be applied in marking out phenomenal from mental presentation (global availability for attention, cognition, and behavioral control). As I hastened to point out, this was an extremely oversimplified fiction from an empirical point of view, because there obviously is more than one kind of attention (e.g., deliberately initiated, focused high-level attention and automatic low-level attention), different styles of thought certainly exist, and the behavioral control exerted by, for example, an animal may turn out to be something entirely different from rationally guided human action control. In chapter 3 I then submitted a more comprehensive set of ten multilevel constraints for discussion. Both of these examples may serve as an illustration of the principle of theory relativity. What appears as simple under my provisional, theory-driven attempt at describing the phenomenal landscape of what we called "qualia" in the past may appear as a host of different things if we try to satisfy alternative sets of constraints proposed by data-driven, bottom-up approaches. One man's primitive is another man's high-level theoretical entity. I have tried to respond to this problem by formulating multilevel constraints containing a large number of placeholders that can be filled in by other disciplines. Therefore, to give another concrete example, the candidates I have offered as potential primitives of chromatic vision-Lewis qualia, Raffman qualia, and Metzinger qualia as enriched by constraints I to I0-are entirely relative to my way of describing the domain of conscious experience, to my own set of constraints. The advantage is that my notion of "phenomenal simplicity" is now domain-specific, can be corrected, and continuously differentiated by, say, additional data from psychophysics. Theory relativity is not an epistemological tragedy-it can be defused by making the set of constraints one wants to find solutions for a flexible and evolvable set.

What is the minimal set of constraints that have to be satisfied for conscious experience to emerge at all? For instance, could qualia exist without the global property of consciousness, or is a qualia free form of consciousness conceivable?

A large part of the answer to this question was already contained in the very first answer to the very first question. I believe that the minimal philosophical intuition behind the concept of consciousness is that of the appearance of a reality, here and now. This

amounts to the demand that constraints 2, 3, and 7 have to be satisfied. Personally, however, I think that the most interesting target phenomenon is full-blown cognitive subjectivity, including its social correlates, in terms of satisfying all the constraints sketched in this book. In passing, let me briefly point out how there are even stronger research targets, and that they may interestingly be embodied in our ancient, traditional notions of consciousness. Stronger and centuries-old concepts, like the Greek syneidesis, or its Latin successor conscientia, involving not only an integration of representational space but concomitant higher-order cognition plus moral subjectivity, would inevitably háve to involve all constraints, plus a theory of intentional content, plus a theory of normative mental judgment (Metzinger and Schumacher I999). In order to understand why conscientia is also higherorder knowledge, the current approach would have to be considerably extended toward an integrated theory of phenomenal and intentional content as such. Second, it must be noted, in discussing social correlates, we have so far only proceeded toward an extremely sketchy account of rational intersubjectivity. Moral intersubjectivity and its neurophenomenological conditions of possibility are quite another matter. Conscientia in terms of phenomenally mediated conscience is still beyond us. Therefore, consciousness research, at least when oriented toward those traditional epistemic goals, eventually will have to face issues much more comprehensive than those of the necessary and sufficient conditions determining the appearance of phenomenal content as such. But obviously it is much too early for this.

Could "qualia-free" consciousness exist? Could there be a heavenly place for disembodied mathematicians, dealing with abstract objects only, lacking any forro of sensory perception? On the representational level of description this may seem to be a conceptual possibility. There could be integrated, transparent reality-models held in working memory, involving no forro of presentational content as such. But what about the earlier processing stages? What about the standard causal linkage? If we demand that the stimulus source to which presentational content is strictly correlated is an extradermal source, then even normal dreams could count as such a qualia-free forro of consciousness. Obviously, this would be far-fetched from a phenomenological perspective, because representational atoms certainly do exist in the way in which the dreaming brain emulates actual sensory perception. In any case, the background assumptions of the current approach do not permit consciousness without sensory elements. In particular, it would be hard to translate the representational constraints of our celestial mathematicians into functional level properties. Of course, functionalism does not necessarily imply token physicalism. Functional properties are ontologically neutral in that they could be realized on nonphysical hardware. Angels, qualia-free mathematicians, and the like might be some sort of Turing machine, entirely realized on "angel stuff," but a closer look reveals considerable diffi culties concerning a detailed specification of the causal linkages underlying input and output. Even

on the representational level itself it may be hard to see how the presenta tionality constraint could actually be satisfied. Constraint 2 demands that all conscious content is content de nunc. Abstract objects and other elements of the mathematician's universe, however, are timeless entities: They have no temporal properties. So how could, they, as such, be consciously experienced as being present now?

What is phenomenal selfhood? What, precisely, is the nonconceptual sense of ownership going along with the phenomenal experience of selfhood, or of "being someone"?

Phenomenal selfhood is a representational phenomenon. As explained aboye, any system operating under a transparent self-model will instantiate a phenomenal self if it satisfies the set of constraints we deem minimally sufficient for conscious experience as such. The sense of ownership is a representational property as well. It consists in a content being integrated into the current PSM. Whatever is a part of our conscious self-model is something that we have appropriated. Appropriation in this sense is based not on rule-based inferences or operations on syntactically structured symbol tokens, but on an entirely sub-personal, automatic process of dynarnic self-organization. It is subsymbolic ownership. No sense of cognitive agency is involved.

If the PSM into which information is currently integrated is of a nonhallucinatory kind, then there is not only phenomenal ownership and appropriation but also a corresponding type of intentional content. The system then represents certain aspects of reality as being parts of itself, and it does so correctly. What it achieves is not only self-experience but self-knowledge. This is important, because it brings about a major shift in correlated functional properties as well. The system as a whole now establishes a causal relationship to certain aspects of reality by integrating them into a globally available representation of itself as a whole. Representational ownership goes along with functional integration. The more self-knowledge a system acquires, the more it is able to functionally appropriate the relevant aspects of itself. The phenomenal variant of self-knowledge is of particular relevante here, because it helps to monitor newly discovered aspects of oneself and because it enhances the functional profile of representationally appropriated aspects of oneself, for example, by making them globally available for focal attention, selective cognitive processing, and flexible behavioral control. It is this ongoing process of dynamic, globally available, and functionally active self-integration, which is mirrored on the phenomenal level. It is mirrored as the transparent experience of ownership, as agency, as dramatically enhanced self-control, as flexibility, selectivity, and autonomy-in short, as the ultrarealistic experience of being someone. And again it is important to note that functional appropriation on the subsymbolic level is likely to come in many different degrees. The

actual correlation strength between different parts of the self-model in the brain may vary greatly, for instance, during the process in which a new content is acquired as a newly learned property of the system. In pathological configurations like schizophrenia it may also vary as a function of a process in which representational content is actually lost. Representationally losing a part of the self-model inevitably means losing information, which in turn is equivalent to losing a computational resource for self-control.

It is interesting to note how such fluctuations in internal coherence are also mirrored on the phenomenological level. Phenomenal appropriation is characterized by a complex and ever-changing landscape. The phenomenal sense of ownership is not one simple property, a rigid characteristic of our inner life. For example, it has a subtle temporal profile; it is context-sensitive and it unfolds over time. Taking the phenomenology of ownership seriously means to do justice to the fact that even in normal life the degree to which, for example, you prereflexively experience a certain property of your body or of your emotional personality as your own, or rather as something externally caused and not really belonging to your "true identity," is highly variable. It changes, for instance, as you grow older. And, as philosophers and scientists alike know, the degree to which you actually experience a certain thought or argument as your own certainly depends on the response of your social environment. If the response to certain parts of your cognitive self is exceedingly positive and gratifying, the chances are high that you will experience it as a deeply original part of yourself, as something that always belonged to you, and as something that never was appropriated from anywhere else.

How is the experience of agency related to the experience of ownership? Can both forros of phenomenal content be dissociated?

An agent is a system that has a certain degree of selective and flexible motor control, and that has, in the sense just sketched, representationally and functionally appropriated the underlying selection processes by integrating them into its self-model. Consciously experienced agency appears only if these selection processes are part of the PSM. As explained in more detail in sections 6.4.5 and 6.5.3, an essential part of this process is the activation of a volitional PMIR: a representation of "the self in the act of deliberating," integrating the self-model with certain internally simulated actions or goal states. This, then, is a phenomenal representation of a practical intentionality relation, because it transparently represents the system as standing in a certain relation to a set of possible actions or a specific goal state, and it underlies the experience of agency.

Are there phenomenological situations in which there is agency without ownership? Im a more general sense, it is plausible to assume, for example, that patients suffering frorlt akinetic mutism do not experience themselves as agents while still possessing a very basit

conscious sense of ownership of their body as such. As they are awake and exhibit an ori, enting reflex, they are likely to have an integrated phenomenal body image. Everything integrated into this body image exhibits the phenomenal property of "mineness," of non-conceptual ownership. On the other hand, given the reports available, it seems as if thesq patients never constitute a volitional PMIR (they don't act) or a cognitive PMIR (because they are not so much "imprisoned in a jail of immobility," but rather don't have a mirad at all; see Damasio I994, p. 73; I999, p. I0I ff.), and an attentional PMIR, fleeting and instable, can only be triggered from the outside. So there is bodily ownership without attentional, cognitive, or motor agency. However, the remaining question here is if o ership and agency are really dissociated in the same domain. Schizophrenics experiencing thought insertion and introspective alienation may present us with a more specific case.

Phenomenologically, they experience cognitive agency: specific, conscious thoughts are being selected and forced into their own minds, into what I have termed the opaque partition of their PSM (see section 6.4.2). Phenomenologically, there is a cognitive agent-someone else who is thinking or sending these thoughts. That is, the causal history of these states is phenomenally modeled as having an external origin. They are caused by an agent. This agent is not identical to the subject of experience, however. Second, inserted thoughts are certainly owned, in the sense of now being a part of the schizophrenic's inner world, of her PSM, because they are inserted into what phenomenologically she still experiences as her own mind. So there arguably is third-person agency plus ownership in one and the same domain, in the domain of cognitive representational contents. The neurophenomenology of schizophrenia shows how for one and the same individual thought ownership and agency can be dissociated.

The next logical step to take consists in asking if, in one and the same domain of content, and in any type of real-world neurophenomenological configuration, first-person agency ever does coexist with a lack of ownership. Are there any cases in which the phenomenal self is modeled as a bodily, attentional, or cognitive agent, but in which nonconceptual ownership does not exist? Let us look at some potential examples: Can you have a transparent self-model of yourself as actively generating a certain thought without then owning this thought? Can you experience yourself as actively and focally attending to the perceived weight of the book that you are now holding in your hands without automatically owning the ongoing act of attending? Is it possible to deliberately initiate a bodily movement without owning the proprioceptive feedback, without phenomenally experiencing the actual movement as that of your own body? It is very tempting to say that here-if anywhere at all-we are actually confronted with phenomenal necessity, with something lawlike, with an essential connection between two elements holding across all possible cases. Agency, the representation of subpersonal selection processes on the level of the PSM, is a prime candidate for the conceptual essence of phenomenal ownership, for

phenomenal selfhood, and for the deepest origin of subjectivity, simply because the two elements are so strongly correlated. Is this a domain-specific neurophenomenological "law of selfconsciousness?" Is the constitution of a phenomenal self, of consciously experienced ownership causally tied to the mechanism by which systems like ourselves functionally appropriate the subpersonal processes selecting target objects for attention, cognition, and action?

The answer may be yes-but we must not forget how it will only hold for healthy human beings. First-person agency is certainly a sufficient condition for ownership in most nonpathological configurations. However, once again expanding our explanatory target domain, the initial example of akinetic mutism seems to show that it is not a necessary condition: You can phenomenally own your body without being an agent. Our best current control-theoretic models for the delusions of control arising in schizophrenia (e.g., Frith et al. 2000) give us a detailed understanding of how an agent can consciously own his body, and his self-caused bodily motions, without being able to phenomenally appropriate the selection and initiation process leading to them. As a result, he owns a body that feels like a remote-controlled puppet. On the contrary, ownership seems to be necessary for agency. Personal-level, conscious selection processes always operate on elements represented through the PSM, and these elements have become parts of the PSM through subpersonal, unconscious mechanisms of integration which themselves cannot be phenomenally owned. And this is one way in which subjectivity is anchored in the objective world: It necessarily depends on subpersonal, unconscious information processing.

Can phenomenal selfhood be instantiated without qualia? Is embodiment necessary for selfhood?

As we saw in section 2.4. qualia (Le., presentational content satisfying constraints 2, 3, and 7) are representationally atomic. Their atomicity is always relative to a readout mechanism (e.g., attention vs. cognition), and to the set of constraints sufficient for phenomenality for which we have opted on theoretical grounds. The question now is if we can conceive of a transparent PSM that is not made out of representational atoms. This would have to be a self-model which, although itself a representational structure active in some conscious system's information-processing mechanism, would not possess any atoros out of which it is constituted, for example, by binding them into an integrated gestalt-like whole. Are phenomenal primitives necessary for the special case of self-consciousness?

I can imagine two situations in which they are not. First, there could be an integrated self-model but no readout mechanism that creates primitives through its limited resolving power. For instance, primitive organisms could have a very simple self-model possessing no subformats whatsoever and no attentional or cognitive mechanisms by which, through

metarepresentation, they could create such special subregions in their phenomenal self. Their low-level integration mechanisms would simply make them self-conscious, without turning them into attentional or cognitive subjects. They could not attend to or think about themselves. There would never be a second-order PMIR directed at a first-order selfmodel. Yet they could be self-conscious. If their little PSM is maximally simple and does not change over time-that is, if it does not satisfy constraints 4 and 5-then it could not actually exhibit any introspectively discriminable atoros or subregions. There would be no individual and distinguishable sensations, and no modalities. In a way it would be one singular self-quale or one homogeneous self-presentatum, but its singularity and its sil'" plicity would not be equivalent to the representational atomicity that can be created' by a top-down mechanism such as attention. It would be an integrated self-model resulto entirely from the dynamical self-organization of unconscious bottom-up mecharais

Maybe such a phenomenal self even resembles the one a human patient suffering from akinetic mutism is left with.

Then there is a second possibility. We can certainly imagine truly introspective systems possessing a second-order PMIR directed at a first-order self-model. For example, they could attend to the content of their self-model. However, if this content would change in a strictly continuous manner, if it were grainless on the only level of organization on which attentional processing can operate, then there would be no self-representational atoms as well. No discernible internal boundaries would exist. The content of such a self-model would smoothly change, and blend from one form of, for example, interoceptive content into another in a way that would make it impossible to introspectively3 discrimínate steps within this process. Phenomenologically, there would be a differentiated and changing, yet atomless phenomenal seIf. However, it could still be a presentational self in possessing a strongly stimulus-correlated component, functionally anchored in the physical brain.

Very interestingly, this second conceptual possibility may actually be a good way of describing human self-awareness on the proprioceptive level, on the level of gut feelings, subtle background emotions, or motion perception. We are holistically embodied beings. This is to say that there likely are important and interesting layers in the human PSM not even exhibiting anything in terms of Lewis qualia or Raffman qualia (see section 2.4), layers which are not only inaccessible to categorical perception and gestalt-forming grouping operations caused by attentional processing but which form a continuous multimodal mélange. Phenomenologically, isn't it true that there are aspects of the bodily and emotional self which are not only ineffable but-pardon the metaphor-so liquid that there is no way of holding them, fixating them (even for a brief period of time), or "zooming in" on them? You can only be them. If this phenomenological point is not misguided, it might be exactly those levels of phenomenal content in which the mind is closest to its body. Taking the

phenomenology of embodiment seriously may help in discovering the microfunctional level of description on which we can eventually give up the vehicle-content distinction for self-consciousness. One might want to term this the "principle of liquid linkage." There would then be a level of phenomenal embodiment, which is below, and more holistic than, anything we could describe using the traditional terminological machinery of "first-order phenomenal properties," "structureless qualia," or "simple sensations." Given the current theory, in standard situations functional embodiment certainly is a prerequisite of phenomenal embodiment and strong first-person phenomena (e.g., see sections 5.4, 6.2.8, and 6.5.3). But I would submit the idea that carefully describing its low-level Phenomenal correlates would entail creating completely new conceptual instruments, instruments which may be extremely hard to develop. However, it may be precisely such instruments that may bring us much closer to a solution of the self-body problem.

What is a phenomenally represented first-person perspective? How-does it contribute to other notions of perspectivalness, for example, to logical or epistemic subjectivity?

The content of a consciously experienced first-person perspective is the content of a PMIR. Phenomenologically subjective states are ether, in a narrow sense, states that currently form the object component of the PMIR (the "focus" of awareness), or, more generally, states that are integrated into an internal model of reality structured by a PMIR. On the representational level of description, a PMIR can have two forms of content: phenomenal content and intentional content. Please recall that I have not given an explicit theory about intentional content in this book, or about what actually makes a mental or phenomenal state a carrier of information, or about the conditions under which it can embody knowledge (e.g., because it can also misrepresent the current state of affairs). Depending on the shape of such a theory of mental representation we could say the following: Epistemic subjectivity is a phenomenon that appears whenever a fact is represented under a PMIR. A PMIR is a functional mode of presentation. If a given PMIR has not only phenomenal but also intentional, that is, representational content, then it transforms the system not only into a subject of experience but finto a subject of knowledge at the same time. This still leaves many options open to us. Such subjective knowledge could be conceptual or nonconceptual, explicit or implicit; it could be a classic mental symbol á la Fodor or a trajectory through some suitable state space á la Churchland. What counts is the contribution the notion of a phenomenal first-person perspective can make to the concept of an epistemic state. Subjective epistemic states are now characterized by an extremely rich sét of constraints (see chapter 3); they are one way-and indeed a very special way-out of countless other ways, to process information, to generate intentional content, to

model reality. An epistemic first-person perspective emerges whenever a phenomenal first-person perspective is not completely empty in terms of intentional content.

Could there be strong epistemic subjectivity that is completely devoid of phenomenal content? There certainly could, if the present theory was false. For instance, the double satisfaction of the globality constraint and the presentationality constraint could yield only necessary, but not sufficient conditions for the appearance of what we termed "minimO consciousness" in section 3.2.II. In order to realize first-person knowledge without first-person phenomenal experience, there would have to be an unconscious system which possesses an integrated model of reality plus a virtual window of presence (constraints 2 atll'd 3), and which at the same time has a system-model (which now is not a PSM) and 0, internal model of the intentionality relation as such (which now is not a PMIR). All these structures, like all unconscious states, would be neither transparent nor opaque, because transparency and opacity are properties of phenomenal states. Because selfho emerges through the phenomenal transparency of the system model, the unconscious could only portray a system in the act of knowing, but never a self in the act of knowing. But how could that be a strong version of epistemic subjectivity? It could at best be only a weak, functional form of internally modeled knowledge acquisition. The act of knowledge acquisition would be represented, but the "perspective" it generated would be no one 's perspective. In addition, neither the world nor the present nor the system's epistemic perspective could be portrayed as real, as actually being present right now. What would the "first-person" component in a concept like "unconscious first-person knowledge" mean in this case? Ultimately, it seems, epistemic subjectivity is anchored in the phenomenal property of selfhood as well. If we ask for a better understanding of epistemic subjectivity, we always ask for a better understanding of the phenomenal self as subject.

Of course, as the general concept of consciousness is still devoid of empirical content today, many absurd scenarios may still strike us as conceivable. But the set of logically possible worlds is not identical with the set of metaphysically possible worlds and as science moves on and fills the notion of consciousness with empirical content, many of these scenarios will gradually become less conceivable. Still, today one may, for instance, ask: Couldn't there be zombie transparency? In order to follow this line of thought, we would have to assume some kind of purely functional, but strictly nonphenomenal, notion of availability or unavailability of earlier processing stages plus the empirically unlikely possibility of an entirely unconscious form of focal attention to which it could be relative (see section 3.2.7). If any kind of introspection is at all possible-and the existente of an unconscious MIR necessarily implies at least introspectioni, that is, some kind of metarepresentational capacity directed at some aspect of the internal model of reality-then there are only two possibilities. Either earlier processing stages for this aspect are not unconsciously

attentionally available or they are. Either they are (nonphenomenally, purely functionally) transparent or they are (nonphenomenally, purely functionally) opaque.

In the first case, the system almost exhibits the minimal degree of constraint satisfaction in order to speak about the phenomenon of "appearance," the phenomenon of consciousness at all (see section 3.2.II). It involves constraints 2 (presentationality), 3 (globality), and 7 (transparency, but only in the weaker, now purely functional variant), that is, the activation of a coherent and functionally transparent world-model within a window of presence. Ex hypothesi it would be unconscious, but it would still represent a world as now being present to this very system while it is simultaneously embedded in and directed at it. Could one say that a reality "appears" to this system or not? In the second case, the reality-model of the system, including the purportedly unconscious MIR, is functionally opaque. Earlier processing stages are globally available for unconscious attention. Could one say that this system is not a naive realist anymore? If it would self-ascribe the property of not being a naive realist to itself, what would fill the subjectargument place in the expressions it formed? I think the central lesson to be learned is that the target property of our investigation is the property of a knowing phenomenal self and how it would make little sense to claim of an unconscious system that "it" had overcome naive realism.

This leaves us with two interesting borderline cases. First, the empirically unlikely clase -of systems just described, of course, is conceivable from a purely logical point of view:

There could be systems satisfying at least constraints 2, 3, and 6 without strongly satisfying constraint 7, the transparency constraint. Let us say they would have a model of the present, a world-model, a nonphenomenally transparent system-model, and an accordingly impoverished model of the first-person perspective as such. All of these models would be completely unconscious. Of course, such a state could be an epistemic state too; antd arguably we could also opt for it to be called epistemically subjective without being called phenomenally subjective. The system could have knowledge, because it could potentially correctly represent facts under an unconscious MIR. This now seems to be entirely a point of terminological convention. The present theory would have to describe it as unconscious, because the transparency constraint was interpreted as a necessary condition in the conceptual ascription of phenomenality-but please remember that at the same time I have always warned against any attempts to draw absolute lines in a domain as complex as that of conscious experience. There could possibly be a conscious mirror image of the same configuration.

For the case of global opacity, and in the conscious mirror image, the expected phenomenology of this system class would involve neither selfhood nor any forro of naive realism about subject-object relations, and it could probably be best described as a mixture of an enduring lucid dream (see section 7.2.5) and a prolonged mystical experience of ego

dissolution (or "system consciousness," to use our new conceptual tool). It would also be nonsubjective in that, epistemically, it could constitute selfless knowledge, that is, an at least partially correct representation of reality, but again there would be no representation of the subject as self. If it would self-ascribe phenomenal properties, what would fill thes subject-argument place in the expressions it formed? Many people would certainly opt,for a terminological convention describing this type of system as unconscious. The "subject as self" clearly seems to be the intuitive target property we wanted to understand at the outset. This is why I would at the same time not count any possible scenarios of n~ nomenal epistemic states as subjective in any interesting sense. There is no entity wlí'eb is subjected to the evolution of these states. Therefore, coming back to our initial gwCStion, the contribution the phenomenal first-person perspective makes to epistemic and logical notions of subjectivity is that they are inevitably anchored in an implicit notion Of phenomenal selfhood. The second part of our answer is a note of caution: We may haYe to make domain-specific revisions in our catalogue of constraints. But this is good news, because this is precisely what this catalogue was introduced for.

There is a second possible borderline case. A system could fail to satisfy the transparency constraint simply because it had no attentional processing mechanisms. It follows that its representational states would be neither phenomenally opaque nor phenomenally transparent. It could have an internal model of the present, a world-model, a self-model, and some sort of nonattentional model of the first-person perspective as such, but neither content properties nor vehicle properties would be available for introspective attention. By the definition introduced in section 3.2.II this system would not be minimally conscious. However, please note how this scenario does not exclude the possibility that this system possesses some sort of cognitive introspection. If something doesn't have an attentional PMIR, at least from the conceptual point of view this doesn't exclude the possible existente of a cognitive PMIR or a volitional PMIR.

Imagine nonbiological information processing characterized by a truly classicist architecture: Only GOFAI-type symbolic representations would exist in its internal ecology of epistemic states and all there was would be rule-based operations on syntactically specified tokens-and let us just grant that it could under these conditions actually have some kind of knowledge about the world and itself. Our artificial demon would have a symbolic world-model, a symbolic self-model, and a symbolic model of the intentionality relation as well. Their content would be cognitively available to the classicist demon, as it could in some sense think about it, by forming concept-like mental structures. In particular, it would only have an impoverished and purely cognitive first-person perspective, but not an attentional first-person perspective. In generating it, it could link some demonstrative self-symbol to, say, an object-symbol, forming a more complex internal expression of the sort (THIS system is currently being affected by a perceptual object belonging to category X in manner

Y). It could be an epistemic agent, even a superb autonomous agent, a truly cognitive robot marching through the world, while continuously extracting information through its sensors and generating a nonbiological kind of purely symbolic knowledge. If its model of reality were a good representation of reality, then it would have intentional content. But would it have phenomenal content? It seems easy to dream up a possible world in which it would be an epistemic subject, but not an attentional subject. Again, this seems to be a matter of terminology. The present theory excludes the possibility that this system could be a conscious subject, because, for reasons of empirical and phenomenological plausibility, the SMT puts a strong emphasis on subsymbolic information processing. It is hard to see how something like our classicist demon could have evolved. And in this case, I would like to point out, our intuitions in favor of the thesis that the artificial demon is definitely not conscious are much stronger. But then again, intuitions only reflect what was phenomenally possible and necessary in our lives and the lives of our ancestors.

Phenomenality comes in theory-relative degrees of constraint satisfaction. And intuitions can be chauvinistic.

Can one have a conscious first-person perspective without having a cónscious self? Can, one have a conscious self without having a conscious first-person perspective?

By definition, there can be no PMIR without a subject component. However, as pointed out in section 6.2.6, it is conceivable that this subject component is entirely opaque, and for this reason would not serve to instantiate the phenomenal property of selfhood. In such a case there could be consciousness, plus a PMIR, but one originating from what phenomenologically would be an opaque system-model only. So the first part of our answer is that on the level of conscious experience there would be no one having the respectiva perspective. We can describe the possibility of this type of neurophenomenological con, figuration in a way that involves no logical contradictions. It is an open empirical question whether the class of systems described by this configuration is functionally possible. It is an open philosophical question if in such cases we would still want to speak of a first-person perspective.

By definition, a PSM is a theoretical entity distinct from a PMIR on all levels of description. Therefore, it should be possible for a PSM to exist without a PMIR. Empirically, examples like the case of akinetic mutism discussed aboye constitute plausible evidence that neurophenomenological configurations of this type actually exist. Philosophically, it is a much more difficult question if the existence of a stable PMIR should be treated as a logical condition for actualized personhood, or if an impoverished, but stable, PSM is enough to ascribe personhood. Fortunately, since many patients suffering from akinetic mutism recover after some time, they certainly all possess the potential to

regain phenomenal subjectivity all by themselves. In this sense they certainly are persons. But given the conceptual instruments now at hand, it easy to see that much more difficult cases may exist. Such cases could be constituted by people still possessing a phenomenal self, in whatever rudimentary way, but having lost all potential of regaining rationality, a cognitive or a volitional PMIR.

In which way does a phenomenal first-person perspective contribute to the emergence of a second-person perspective and to the emergence of a first-person plural perspective? What forms of social cognition are inevitably mediated by phenomenal self-awareness. Which are not?

Neurophenomenologically, a second-person perspective consists of the brain activating a PMIR with the object component being another person or self-as-subject. Neurophenomenologically, a first-person plural perspective corresponds to a PMIR with the subject component being represented as a group of persons or selves. However, this locally reductive approach only helps in understanding the conscious experience of different aspects of intersubjectivity, its phenomenal content (i.e., precisely that aspect which could also be a hallucination taking place in an isolated individual). If we want to understand what the functional foundations of real-world, successfully unfolding social cognition are, then we have to assume a more complex situation: Two individuals currently "coherencing" their PMIRs by mutually making each other their object components while at the same time also consciously representing the fact that the other self-as-subject-as-the-object-Iam-currently-directed-at also has conscious knowledge about what is currently taking place; or a group of individuals "orchestrating" their PMIRs by making a mental representation of their group an element of their individual subject components, again paradigmatically knowing that it is exactly this that is now taking place. In this type of situation, we have not only the conscious experience of an "I-Thou relationship" or of (let's say) group solidarity but also the necessary functional properties underlying social neurophenomenology. A dynamical functional equivalence holding across individuals is established. What forms of social cognition are necessarily mediated by conscious selfrepresentation? We can now give a rather abstract, but straightforward answer: All forms of social cognition that make the additional functional demand for a successful transmission of globally available system-related information on both sides, or in all interacting participants. More concretely, a PSM and a PMIR are needed in all those cases where selectivity and flexibility are important: if social interaction necessarily involves the fast and flexible adaptation of your own behavior to that of another human being (global availability for action control), constant metacognitive monitoring of your own thoughts (global availability for cognition), or a permanent introspective observation of your ongoing emotional responses (global availability for self-directed attentional processing). And, of course, what we have called

the virtual window of presence (constraint 3) is necessary for you and the other to be able to represent each other as continuously interacting at the same time.

Please recall a point touched upon earlier: Coherent representational structures can function as unifying windows allowing parts of a system to communicate as wholes, with causal forces in the environment, or with other functional subcomponents internal to the system. This well-established principie now has an interesting extension into the domain of social environments. A PSM is a representational structure creating a functional window through which a system can interact and communicate with other agents, with other self conscious systems, with those aspects of its environment that are exclusively formed by other systems also acting and internally representing themselves as wholes. Therefore, a PSM is a necessary functional window onto more flexible and selective forms of agent agent interaction. It creates a new macrolevel of information flow and causal interaction.

It must have been at the heart of the process of forming more complex, evolvable, and yet stable societies. The new level of information flow is also a new level of functional granularity for intrasocial self-representation.

This point can now be made clearer if we consider the two examples of specifically social kinds of PMIRs just discussed, namely, the second-person perspective and the first-person plural perspective. Given our new conceptual tools, we can consider the evolution not of individuals, but of groups of conscious systems and say the following: PMIRs of the sort described above-made coherent and orchestrated-are unified functional windows that emerge in groups of biological systems. They are new causal properties. Of course, societies are information-processing and representational systems as well, and although they do not exhibit any mysterious kind of "group consciousness" they certainiy come in degrees of intelligence. They also create self-models. If new functional windows-new units of transindividual representation-appear in such societies, they can in principle greatly increase their overall intelligence and adaptivity (their degree of "selfreflexivity," if readers will permit this metaphor). My proposal is that phenomenally represented I-Thou relationships and the functionally orchestrated conscious representation of the we-as-a-group emerging in groups of biological individuals are in the same way to be understood as a superbly elegant strategy in which parts of the group can now form transient, but stable functional windows through which the group can now causally and informationally interact with itself and other groups. The point is that the essential neurocomputational feature within the brains of individuals in order to achieve this step on more interesting levels of complexity must have been the human PSM.

Finally, a last question concerns the status of phenomenal universals. Can we define a notion of consciousness and subjectivity, which is hardware and species independent? This issue has a distinct philosophical flavor, because it amounts to an attempt to give an analysis of consciousness, the phenomenal self, and the first-person perspective that operates on

the representational and functional levels of description alone, maximizing generality and aiming at a liberation from any kind of physical domain-specificity. Can there be a universal theory of consciousness? Today, we usually put it in other words:

Is artificial subjectivity possible? Could there be nonbiological phenomenal selves?

To actually create a technical model of full-blown, perspectivally organized conscious experience seems to be the ultimate technological utopian dream. It would transpone the evolution of mind onto an entirely new level-not only in terms of the physical properties on which mind now supervenes but also with regard to functional constraints and optlmality conditions operative during its future development from this point onward. It would be a historical phase transition. And indeed, the project of realizing ever-stronger forma of intelligence, of coherent and content-bearing mentality, and possibly even phenomenal seljhood on artificial carrier systems is a fascinating one. But is this at all possible? It certainly is conceivable. But can it happen, given the natural laws governing this universo and the technical resources at hand? Let us here distinguish conceptual, nomological, and technological possibility (Birnbacher I995). If an exhaustive representationalist and functionalist analysis of phenomenal content is possible, then conscious experience will be multirealizable. It may, however, be the case that in our own physical universe there is only one type of hardware-the human brain-which can actually realize precisely the human kind of consciousness, phenomenal selfhood, and perspectivalness. We simply don't know this. What would be needed is a physical substrate possessing a topologically equivalent phase space to the phenomenal state space of human beings. To create an artificial PSM plus PMIR, a consistent functional equivalence would have to be achieved, and on just the right level of granularity. But please note that such a system would not yet have to be intelligent, or even embodied: An appropriately stimulated chunk of nervous tissue could in principle exhibit the right kind of topological or microfunctional equivalente, without possessing any kind of intentional content. As its internal states would neither be grounded in the environment nor in its history in a way that could endow them with meaning, it would not have any form of knowledge. In particular, this chunk of nervous tissue would not possess self-knowledge, but only self-experience.

Cognitive robotics may soon change this situation. Just as Mother Nature first created unconscious forms of information processing and representation, and phenomenal experience only recently, it might plausibly be argued that the second evolution of minds will have to repeat an unconscious bottom-up phase as well. Embodiment, sensorimotor integration, and unconscious self-models will have to come first. On the other hand, one thing seems safe to say: The smooth and reliable type of ultrafine-grained self-presentation based on molecular-level dynamics-which, in human beings, drives the incessant selfstabilizing

activity in the homeostatic system of the brainstem and hypothalamus-will be out of reach for a long time. The subtlety of bodily and emotional selfhood, the qualitative wealth and dynamic elegance of the human variety of having a conscious self, will not be available to any machine for a long time. The reason is that the microfunctional structure of our emotional self-model simply is much too fine-grained, and possibly even mathematically intractable. And, for the same reason, in terms of what is technologically possible, the portability of human PSMs is extremely low. Self-models emerge from elementary forms of bioregulation, from complex chemical and immunological loops-and this is simply something machines don't possess. The time when robots come to have body fluids and something even remotely resembling the complex homeodynamics of the human brain certainly is far distant. Or is it?

The new discipline of hybrid biorobotics may soon change this situation, by taking the hardware from what Mother Nature has to offer. Please remember our more extensive discussion of the strength or weakness concerning conscious systems not-or differently-satisfying constraint II (the adaptivity constraint) in section 3.2.II: the distinction between natural and artificial systems is not an exhaustive and exclusive distinction.

Postbiotic systems fall into neither category. They might be hybrid biorobots using organic, genetically engineered hardware, or semiartificial information-processing systems employing biomorphic architectures. At the same time they could be submitted to a quasi-evolutionary optimization process of individual development and group evolution. If their PSM is actually anchored in biological hardware, things might be different. Presently, we have to admit that both of our questions concerning the nomological and technological possibility of nonbiological consciousness, and of postbiotic PSMs and PMIRs in particular, are simply open. They constitute empirical, not philosophical, issues. However, please also recall how one of the central lessons to be learned in the course of this investigation was that consciousness and self-consciousness are graded phenomena. There are degrees of constraint satisfaction and degrees of phenomenality. There will be degrees of phenomenal selfhood too. Therefore, just as with animals and many primitive organisms surrounding us on this planet, it is rather likely that there will soon be artificial or postbiotic systems possessing simple self-models and weaker forms of conscious experience in our environment. One aspect that these simple, nonbiological subjects will have in common with us is the capacity to suffer.

It is time to lay my cards on the table. There is a philosophical issue that has been neglected and which cannot simply be naturalized by gradually handing it over to the empirical mind sciences. In the end, a theory of mind has to be rationally integrated with normative considerations. Philosophy of mind must be supplemented by moral and eventually even by political philosophy. Aboye, I pointed out that the actual creation of a technical model of a full-blown, perspectivally organized conscious experience seems to be the

ultimate technological Utopian dream. It might be a nightmare too. As a philosopher I am strictly against attempting to realize the Big Technological Dream, but on ethical grounds. Why? Put very simply, we might dramatically increase the amount of suffering, misery, and confusion on the planet. And we might do so without at the same time increasing the amount of pleasure and joy. An even deeper and more general point is that upon more careful inspection it is not at all clear if the biological form of consciousness, as $0 far brought about by evolution on our planet, is a desirable form of experience, an actual good in itself, something that one should simply keep on multiplying without furthcr thought. Let me explain.

Perhaps the theoretical blind spot of current philosophy of mind is the issue of COIt scious suffering. Thousands of pages are being written about color qualia or the contenta of thought, but almost no theoretical work is devoted to ubiquitous phenomenal stateslike physical pain or simple everyday sadness ("subclinical depression"), or to the phenomenal content associated with panic, despair, and melancholy, let alone the conscious expsrience of mortality or of losing one's dignity. There may be deeper evolutionary reaso'us behind this cognitive scotoma, but I am not going to pursue this point here. The ethics issue is of greater relevance. If one dares to take a closer look at the actual phenomenology of biological systems on our planet, the many different kinds of conscious suffering are at least as dominant a feature as are color vision or conscious thought, both of which appeared only very recently. Evolution is not something to be glorified. One wayout of countless others-to look at biological evolution on our planet is as a process that has created an expanding ocean of suffering and confusion where there previously was none. As not only the simple number of individual conscious subjects but also the dimensionality of their phenomenal state spaces is continuously increasing, this ocean is also deepening. Obviously, the process as a whole is something that has not yet ended. We should not accelerate it without need.

As this is not the place to enter into an extended ethical discussion of artificial phenomenality, let me just give two concrete examples. What would you say if someone carne along and said, "Hey, we want to genetically engineer mentally retarded human infants! For reasons of scientific progress we need infants with certain cognitive and emotional deficits in order to study their postnatal psychological development-we urgently need some funding for this important and innovative kind of research!" You would certainly think this was not only an absurd and appalling but also a dangerous idea. It would hopefully not pass any ethics committee in the dernocratic world. However, what today's ethics committees don 't see is how the first machines satisfying a minimally sufficient set of constraints for conscious experience could be just like such mentally retarded infants. They would suffer from all kinds of functional and representational deficits too. But they would now also

subjectively experience those deficits. In addition, they would have no political lobby-no representatives in any ethics committee.

If they had a transparent world-model embedded in a virtual window of presence, then a reality would appear to them. They would be minimally conscious. If, as advanced robots, they even had a stable bodily self-model, then they could feel sensory pain as their own pain, including all the consequences resulting from bad human engineering. But particularly if their postbiotic PSM were actually anchored in biological hardware, things might be much worse. If they had an emotional self-model, then they could truly suffer possibly even in degrees of intensity or qualitative richness that we as their creators cannot imagine, because it is entirely alien to us. If, in addition, they possessed a cognitive self model, they could potentially not only conceive of their bizarre situation but also intellectually suffer from the fact that they never had anything like the "dignity" so important to their creators. They might be able to consciously represent the obvious fact that they are only second-rate subjects, used as exchangeable experimental tools by some other type of self-modeling system, which obviously doesn't know what it is doing and which must have lost control of its own actions long ago.

Can you imagine what it would be like to be such a mentally retarded phenomenal done of the first generation? Alternatively, can you imagine what it would be like to "come to" as a more advanced artificial subject, only to discover that, although possessing a distinct sense of self, you are just a commodity, a scientific tool never created and certainly not to be treated as an end in itself?

A lot more would have to be said at this point. Let me just highlight what seems to be the central issue: Suffering starts on the level of PSMs. You cannot consciously suffer without having a globally available self-model. The PSM is the decisive neurocomputational instrument not only in developing a host of new cognitive and social skills but also in forcing any strongly conscious system to functionally and representationally appropriate its own disintegration, its own failures and internal conflicts. Phenomenal appropriation goes along with functional appropriation. Evolution is not only marvellously efficient but also ruthless and cruel to the individual organism. Pain and any other nonphysical kind of suffering, generally any representational state characterized by a "negative valence" and integrated into the PSM, are now phenomenally owned. Now it inevitably, and transparently, is my own suffering. The melodrama, but also the potential tragedy of the ego both start on the level of transparent self-modeling. Therefore, we should ban all attempts to create (or even risk the creation of) artificial and postbiotic PSMs from serious academic research.

People differ widely in their positive moral intuitions, as well as in their explicit theories about what we should actively strive for. But in terms of a fundamental solidarity of all suffering beings against suffering, something that almost all of us should be able to agree

on is what I will term the "principie of negative utilitarianism": Whatever else our exact ethical commitments and specific positive goals are, we can and should certainly all agree that, in principle, and whenever possible, the overall amount of conscious suffering in all beings capable of conscious suffering should be minimized. I know that it is impossible to give any truly conclusive argument in favor of this principle. And, of course, there exist all kinds of theoretical complications-for example, individual rights, long-terco preferences, and epistemic indeterminacy. But the underlying intuition is something that can be shared by almost everybody: We can all agree that no additional suffering should be created without need. Albert Camus once spoke about the solidarity of all finite beings against death, and in just the same sense there should be a solidarity of all sentient beings capable of suffering against suffering. Out of this solidarity we should not do anything that would increase the overall amount of suffering and confusion in the universe-I0 alone something that highly likely will have this effect right from the beginning.

To put it very carefully, one obvious fact about phenomenal experience as it has developed on our planet until now is that one of its strikingly dominant features is suffering and confusion. Phenomenal experience is not something to be unconditionally glorified-Among many other new properties, biological self-consciousness has brought an enormous amount of misery and confusion into the physical world, an ocean of phenomenal suffering, which simply was not there before. As one of my students once put it: The universe may be a good place for evolution, but not such a good place for individuals. If this is true, individuals with conscious self-models will automatically reflect this fact on the level of their own phenomenal experience. It is hard for beings like us to really face this fact, because this is not the kind of fact Mother Nature wanted us to face. But in theoretically and technologically modeling fascinating phenomena like phenomenal selfhood and the first-person perspective we now have nothing much to go by other than our own, biological form of consciousness-simply because it is the only form of consciousness we can scientifically investigate. We are therefore in great danger of multiplying all its negative aspects on artificial carrier systems before we have understood where all diese negative aspects come from, in exactly what properties of our biological history, of our bodies, and of our brains they are rooted, and if or how they can at all be neutralized. For this reason we should first focus all our energy-in philosophy as well as in the neuroand cognitive sciences-on achieving a deeper understanding of our own consciousness and the structure of our own suffering. We should orient ourselves in accordance with the classical philosophical ideal of self-knowledge and the minimal ethical rule of the minimization of suffering rather than risk the triggering of a second-order evolution of postbiotic minds, which could then slip from our control and eventually contribute to a further increase in the overall amount of suffering in the universe.

Before closing by briefly considering more general and normative issues in the final section, let us stop to ask: What are the most urgent goals for future research? Where do we go from here? The beauty of the current phase of interdisciplinary research lies in the fact that the new image of the conscious mind that is now slowly emerging is the first image in the history of mankind that is anchored in a firm empirical foundation. Therefore, on our way toward a new theory of mind, this is also the first image justifying serious hopes for clearly nameable steps of progress. It is hard to underestimate ¡he relevante of this fact for the old philosophical project of a comprehensive and unified theory of consciousness. What are rational next steps? On the level of neuroscience we should first focus on the minimally sufficient neural correlate for the PSM and the PMIR:

• Which content layers of the PSM covary with which types of neural processing?

• How is a dynamic binding of these layers achieved?

• How are we to imagine the way in which the PSM is anchored in unconscious processes of self-representation? What is the most simple neural structure in the brain that can still be said to represent the system as a whole?

• In terms of neural correlates for the PMIR, what are candidates for object-components in different domains (e.g., in perceptual attention, the selection of motor patterns, or in conscious concept formation)?

• At any given point in time, how is the dynamic binding of subject and object component into a single PMIR achieved?

• What are the unconscious processing stages necessarily preceding the activation of a PMIR?

• What exactly is the minimal set of neurobiological properties that will bring about a conscious self and a consciously experienced first-person perspective in humans?

• How does this set contribute to the set enabling intersubjectivity and social cognition?

On various functional levels of analysis we need more detailed descriptions of the causal and informational fine structure for given physical correlates. On the level of fine-grained functional mapping and computational modeling we need more abstract descriptions of the PSM, as well as of the PMIR:

• What is the functional neuroanatomy of the PSM?

• In terms of an extended teleofunctionalist analysis, is there something like a distinct biological proper function of self-consciousness and phenomenal perspectivity?

• What, precisely, is the computational role of the phenomenal self and the first-person perspective under a more abstract, mathematical description?

• How is this computational role integrated into the behavioral ecology of the system, for example, into sensorimotor loops, into the ongoing generation of more complex motor output, and into other-agent modeling and social cognition?

• Ontogenetic development: We certainly need to know more about the stages and the general developmental trajectory through which individual self-models unfold in individual human beings and other animals.

• Phylogenetic history: If my claim is true that the PSM and the PMIR are "virtual organs" that have been developed in the course of biological evolution, then it must be possible to tell an evolutionary story for individual species and the way in which they developed these organs in order to adapt to their specific inner and outer environment. How did PSMS propagate through biological populations?

There a number of important and largely unresolved conceptual issues as well. TheY begin on the representationalist level of analysis, expanding into phenomenology and ethics. These future goals mainly fall into the field of philosophy. Most pressing may be the relationship between phenomenal and epistemic subjectivity:

• Which aspects of the contents of self-consciousness can be epistemically justified?

• What is the difference between a phenomenal and an epistemic first-person perspectiva - can one exist without the other; in what way do they depend on each other?

• More generally, we urgently need an overarching theory of mental content that explains the relationship between intentional and phenomenal content while at the same time satisfying empirical constraints as they are, for instance, given by the best current theories in connectionist/dynamicist cognitive science. What is needed is an empirically plausible theory of mental content, which is open to future changes.

• On the phenomenological level of analysis new tools have to be developed. As ultimately they are always phenomenological descriptions that function as input for the method of representational analysis, these descriptions have to be optimized beyond the terminologies of classic philosophical phenomenology developed in the tradition of Brentano and Husserl. Even if nothing like "first-person data" in any stronger epistemological or methodological sense exist, the heuristic power of first-person reports has been underestimated. Careful introspective reports, particularly in combination with real-time thirdperson access through neuroimaging, transcranial magnetic stimulation, and so forth, are an important source of information in correlation studies. Therefore, innovative methods for arriving at more precise first-person descriptions of the target phenomenon are of highest relevance.

• The preliminary catalogue of constraints offered in chapter 3 needs to be critically assessed, continuously differentiated, and expanded. Further domains and more finegrained levels of description have to be added.

• Normative issues and cultural ramifications have to become topics of a permanent discussion accompanying progress in the cognitive neuroscience of consciousness and self-hood. Because it has obvious political aspects, this discussion can not exclusively remain an expert discussion, but eventually has to include the general public. If the current proposal

points in the right direction, then it is an obvious fact that we are facing a major shift in our general image of humankind and that a host of new ethical issues will eventually result. In the face of rising time pressure an important task of academic philosophy lies in offering a service to society as a whole by initiating critical and rational debates on these issues.

8.3 Being No One

What does it mean to be someone? "Being someone" is not a well-defined technical term, neither in philosophy nor in any other discipline. Therefore, it simultaneously means many different things to different people. We all use the idea of "being someone" in many different ways, and in many different contexts-as citizens of a state or as psychological laymen, in ethical and political discourse, or even in religious matters. As this is only a book about consciousness, the phenomenal self, and the first-person perspective, I have been mostly interested in the phenomenological aspects of this question: What exactly does it mean to have the conscious experience of being someone? In this limited sense, the folk-phenomenological notion of "being someone" denotes a phenomenal property like many others, a property like the scent of mixed amber and sandalwood or the gustatory experience of cinnamon, a property like the emotional experience of elation, or the sense of surprise going along with a sudden cognitive insight. It is just a way of experiencing reality: currently, you are someone. What makes consciously experienced selfhood special, and different from all the other forms of experiential content, is the fact that-in nonpathological standard situations and in beings like ourselves-it is highly invariant. It is always there.

This phenomenally transparent representation of invariance and continuity constitutes the intuitions that underlie many traditional philosophical fallacies concerning the existence of selves as process-independent individual entities, as ontological substances that could in principie exist all by themselves, and as mysteriously unchanging essences that generate a sharp transtemporal identity for persons. But at the end of this investigation we can clearly see how individuality (in terms of simplicity and indivisibility), substantiality (in terms of ontological autonomy), and essentiality (in terms of transtemporal sameness) are not properties of selves at all. At best, they are folk-phenomenological constructs, inadequately described conscious simulations of individuality, substantiality, and essentiality. And in this sense we truly are no one. We now arrive at a maximally simple metaphysical position with regard to selves: No such things as selves exist in the world. At least their existence does not have to be presupposed in any rational and truly explanatory theory. Metaphysically speaking, what we called "the self' in the past is neither an individual nor a substance, but the content of a transparent PSM. There is no unchanging essence, but a complex self-representational process that can be interestingly described on many different levels of analysis at the same time. For ontological purposes, "self' can therefore be substituted by

"PSM." However, this first reading of the concept of "being no one" is only an answer to the crude traditional metaphysics of selfhood, and I think as such it is a rather trivial one.

On a somewhat deeper level the question arises if the dominant structural characteristic of our phenomenal space-the fact that it almost inevitably satisfies constraint 6, the perspectivalness constraint-makes us constitutionally unable to see certain obvious truths. Could the fact that we always operate not only under a transparent PSM but also under a PMIR impede epistemic progress? There is one obvious field of research at which this question is aimed: the now strongly expanding domain of the mind sciencesscientific psychology, cognitive neuroscience, Al and robotics, philosophy of mind, and the like. More specifically, could it be that the conscious experience of being someone itself hinders growth of knowledge in these disciplines, by making certain theoretical positions or solutions of problems look utterly implausible, dangerously provocative, absurdly humiliating, or simply inconceivable to beings like ourselves? A lot of today's physics, for example, describes the world in a way that is extremely counterintuitive, and certainly hard to conceive of. Yet most of us believe that these theories are among the best mankind has so far created. Basically, we trust those physicists. In the mind sciences things are different, and in an interesting way.

Take as an example the sketch of an interdisciplinary, representationalist theory of consciousness, the phenomenal self, and the first-person perspective I have offered in this book. Even if you should think that at least some of the ideas involved are potentially worthy of discussion, you could never really believe that the SMT, the self-model theory of subjectivity, actually is true. You cannot believe in it. Take what may be the central idea, the idea that metaphysically speaking no such things as selves exist in the world; that the conscious experience of selfhood is brought about by the phenomenal transparency of the system-model; and that what philosophers call the epistemic irreducibility of conscious experience-the fact that it is tied to a first-person perspective-can be exhaustively analyzed as a representational phenomenon, which in the future will likely be fully explained on functional and neurobiological levels of description. You cannot believe in the truth of this idea. "Being convinced," like smelling mixed amber and sandalwood or being someone, is here interpreted as a phenomenal property. But for the current theory you cannot in principie have that property, because phenomenally simulating the truth of the SMT would involve a cognitively lucid, nonpathological way of dissolving your sense of self. It would involve being convinced and phenomenally being no one at the same time.

My second conclusion in this final section therefore is that the SMT is a theory of which you cannot be convinced, in principie. I would also claim that this fact is the true essence and the deepest core of what we actually mean when speaking about the "puzzle"- or sometimes even about the "mystery"-of consciousness. Furthermore, this second conclusion is another possible answer to the question which many readers may have silently

been asking themselves for quite a while: Why is the title of this book Being No One? After all, isn't it precisely a book about the neurophenomenological conditions of personhood, a book that tells a new story about what it means to Be Someone? The problem is this: If the current story is true, there is no way in which it could be intuitively true. It could never feel true, because it creates a dilemma. There seem to be two alternatives: Either you see it as actually describing a set of possibilities that may be nomologically likely (Le., empirically plausible) and conceptually coherent (i.e., philosophically plausible) at the same time. Then you cannot be convinced. Call this the "scientific horn of the dilemma." You cannot be convinced, because the idea that there are no such things as selves -including your own self- in the world remains strictly counterintuitive, a phenomenal impossibility. Now you might turn to the other alternative. Let us call this the "spiritual horn of the dilemma." You might change your global phenomenal model of the world in a way that makes it a possibility. For instance, you could do so by making it a phenomenal reality, that is, by developing a stable and cognitively lucid state of consciousness that does not satisfy constraint 6. Phenomenal selfhood would not be instantiated. Your new neurophenomenological configuration would then correspond to what was earlier termed "system consciousness," namely, a phenomenally nonsubjective state of consciousness. In this case you could not truthfully form the corresponding I*-sentences (see section 6.4.4), and therefore you could not even self-ascribe your new neurophenomenological configuration to yourself. In this case, you could not be convinced of the truth of the SMT, in principie. In conclusion, no one can be convinced of the current theory. And this is another one of the reasons this book has the title it has: "Being no one" in this sense describes an epistemological stance we would have to take toward our own minds in scientifically and philosophically investigating them, an attitude that is necessary to really solve the puzzle of consciousness at a deeper and more comprehensive level, an attitude of research that integrates first-person and third-person approaches in a new way and that, perhaps unfortunately, appears to be strictly impossible and absolutely necessary at the same time. It goes beyond the classic research strategy of methodological solipsism in cognitive science in a new way, because it acknowledges the peed for a shift in perspective that we could call "methodological nemocentrism."

Is all of this a problem? Yes and no. It is a problem if-as opposed to other, for example, physical, theories about the nature of reality-we impose the additional constraint of intuitive plausibility on theories of consciousness, the phenomenal self, and the first-person perspective. It certainly is an absurd claim that simply listening to a theoretical description of the underlying causal reality should create the respective form of phenomenal content in our minds (recall Frank Jackson's Mary). One and the same fact can be given via two different modes of presentation. I would submit that a PMIR may just be such a mode of presentation. The fact that, in standard situations, you are a single and unified physical system operating in its behavioral space under a functionally centered model of reality is

made globally available through a highly specific phenomenal mode of presentation. This is an entirely new epistemic possibility, which, however, does not entail the corresponding metaphysical possibility. There are no new and irreducible phenomenal facts-all there is is a rather complex new way of accessing an internal physical fact under a phenomenal model, under the PMIR mode of presentation.

The phenomenal first-person perspectiva described in chapter 6 is just this mode of' presentation. For the more analytically inclined, we might even call it an "indexical ego mode of presentation."[6] But what is the fact that is given under a PMIR? Strictly speaking, the fact presented is that there currently is a certain brain state, the state on which the PMIR locally supervenes. This brain state can additionally be given under a different mode of presentation, for instance, one involving theories developed by the cognitive neurosciences. The same fact would then also be given under a nonphenomenal, third-person, propositionally structured description. And, of course, it is absurd to demand that reading this description as such could by sheer magic turn you (or a selfless machine) into a specific phenomenal self, tied to a specific phenomenal first-person perspective. Actually implementing a computational model of this theory, however, might be a different matter. So the radically counterintuitive nature of the SMT only poses a problem if we want to extend the usual criteria for the goodness of a theory (such as logical coherence, parsimony, predictive power, etc.) by additionally demanding that it be phenomenally possible. As you may recall, a theory is phenomenally possible relative to a given class of representational systems if and only if these systems are able to emulate its ontology. The selfless metaphysics of the SMT is not an ontology human beings can emulate. As such, this is not a problem, just as it is not a problem that we are unable to consciously emulate the ontology of quantum chromodynamics. The yet deeper question lurking in the background, of course, is if we would ever want lo emulate-or even instantiate-this kind of ontology. "Being no one," therefore, could not only refer to the serious and sustained theoretical effort of thinking the unthinkable but also of the ideal of phenomenally living it.

6. I will not go into analytic details at this point, but just inform my readers that Albert Newen (I997) has introduced the idea of an "indexical ego-mode of presentation," which may be closely related to the more general idea I am sketching here. For instance, Newen writes: "Even though the ego-mode of presentation is not based on any identification, I claim that it nevertheless is a real cognitive structure: a representation that relates one to oneself. To characterize this cognitive structure we have to introduce the distinction between object representations and subject representations.... A mode of presentation is subject representational if it constitutes a mental representation that, first, one could not have if the object the thinker is related to does not exist and, second, that does not allow for misidentifications" (Newen I997, p. I27). I propuse that the PSM

and the PMIR are just the "real cognitive structures" posited by Newen. However, as we have seen in chapter 7, consistent misidentification of otherwise rational subjects actually does occur. Also, given the current theory, it cannot be assumed that there are possible information-processing systems which, as opposed to the cases of Christina and tan Waterman discussed in the same chapter, have never had any proprioceptive information to construct their fundamental bodily self-model and are still able to have thoughts de se. But Newen certainly makes a good point when writing, "Neither the kind of information nor the way the information is acquired, but rather the way the information is handled in a information-processing system is the essential feature that makes it subject representational.... the self is a person who is related to himself/herself in the ego-mode of presentation, i.e., this person has a special repository for indexical information that plays a characteristic role in perception, action, and thinking" (p. I28f). The PSM, in particular its self-presentational layer described in chapters 5 and 6, is this "special repository" for indexical information. In order to anchor the philosophy of self-consciousness in scientific, third-person approaches to the mind via empirical constraints, it is therefore of great importance not to stop here, but to begin describing the representational deep structure and the functional architecture of this neurocomputational tool.

In closing, let us now once again return to our original question of what it could to be no one. The third potential reading I want to explicitly mention relates to the ethics of consciousness: Do we want to phenomenally emulate the ontology of our own scientific theories about the mind? Do we want to instantiate them? My third interim conclusion in this final section is that the cognitive neuroscience of self-consciousness will soon confront us with an extremely interesting set of normative challenges. Some of them are obvious and rather concrete practical issues like, for example, defining an applied ethics for medical neurotechnology, for animal experiments, or the question of rejecting military funding in consciousness research. But some of them possess an even more distinct philosophical flavor, because they are much deeper and of a more general type. Unfortunately an in-depth discussion of such wider normative issues clearly is outside the scope of' this work (but see Metzinger 2000b, p. 6ff.). However, let us at least take a brief look at some examples.

As we have already seen, there is more than one answer to the question of why this book has the title it has. If it is true that we are neurophenomenological cavemen, then it is also true that mankind is still in a prehistoric stage-not in terms of theoretical knowledge and technology, but in terms of phenomenological knowledge and technology. One more general question is if, in the long run, we want to use our new insights into the nature of consciousness, the phenomenal self, and the first-person perspective to change our own minds. Is it better to be someone or is it better to be no one? Is the current neurophenomenological configuration of Homo sapiens really a good in itself? Is this really something we want to perpetuate and multiply indefinitely? Or should we start to think a improving our conscious model of reality, particularly our PSM? Put crudely, we have

better theories and we have better computers-why shouldn't we have better phenomenal selves as well?

In chapter 3 we made an attempt to describe the maximal as well as the minimal degree of constraint satisfaction for subjective experience to occur. Interestingly, one can now also define a notion of optimal constraint satisfaction: If it is true that phenomenal experience comes in many different grades and that human beings possibly possess the highest degree of conscious awareness (at least relative to the preliminary catalogue discused above), then it is only natural to conclude that human beings could also possess a higher degree of consciousness. There is nothing mysterious about this conclusion, which can be formulated in a conceptually clear way: A stronger form of phenomenality simply comes about by a given class of systems satisfying new and additional constraints. Or, as we might decide on normative grounds, less could be more. Of course, there could be more sets of constraints as well, in extraterrestrial beings, in conscious machines, or possibly even in some animals on our planet. Such systems might simply have a very different form of phenomenal experience altogether by satisfying a rather distinct set of constraints, one only loosely overlapping with the one sketched here. The space of possible phenomenal minds is vast. Yet it is interesting to pose the following question: What could additional or different constraints for ourselves actually be?

Normative neurophenomenological considerations could yield such additional constraints. For example, they could do so in terms of maximizing intelligence or minimizing suffering in human beings. Another idea, already alluded to above, and slightly more complex, is to assimilate the implicit ontology underlying our phenomenal model of reality into the ontology of our scientific theories. One might carefully investigate the normative ideal of slowly developing a gradual convergence between human neurophenomenology and the metaphysics implied by our best objective theories about the deeper structure of physical reality. Call this notion "first-person-third-person convergence." A third logical possibility lies in that we could also opt for decreasing the degree of constraint satisfaction for one or more of our already existing constraints. We could, for example, choose to decrease phenomenal transparency. This candidate for a normative orientation-call it "minimization of transparency"-would consist in making the fundamentally representational character of conscious experience globally available. We could attempt to make more information about earlier processing stages available for introspective attention, thereby also gradually making more and more layers in our own selfmodel phenomenally opaque. This type of strategy would certainly create an additional computational load for attentional systems in the brain, but it could at the same time serve to weaken the naive-realistic self-misunderstanding characterizing our present state of consciousness.

So much for first examples. Of course, the number of options open to us is much larger than the three proposals sketched above-from a purely theoretical perspective it is as

vast as the space of possible minds itself, although in present-day human beings it is much smaller due to the contigent neurofunctional constraints resulting from the physical structure of our brain. For all these proposals, the underlying principle would always consist in combining an ongoing scientific discussion of our actual constraint landscape with a normative discussion of what an optimal constraint landscape for human beings could be. In doing so we might, perhaps, eventually arrive at new and more precise answers to ancient philosophical questions like what a good life is and how we can suffer less, about how we can be more intelligent or, more generally, how we can become bearers of a stronger form of conscious experience.

This may also be the point where old-fashioned philosophy reenters the stage. In terms of specific normative aspects concerning potential future changes in the PSM itself, one could, for example, discuss the maximization of its internal coherence. Perhaps (if in a hedonist mood) we could simply set this goal as relative to ever-higher intensities of pleasant self-presentational content: How much physical pleasure can you experience without going insane? How can you use scientific knowledge to optimize sensory stimulation without forcing the self-model to disintegrate? Then there is a related, but already slightly different interpretation of the coherence ideal: The classical notion of "virtue" can now be interestingly reinterpreted, namely, in terms of increasing the intemal and social consistency of the self-model, for example, in terms of functionally integrating cognitive insight, emotional self-modeling, and actual behavioral profile. Traditional notions like "intellectual integrity" and "moral integrity" now suddenly possess new and obvious interpretations, namely, in terms of a person having a highly consistent self-model. Ethical behavior may simply be the most direct way of maximizing the internal coherence of the self-model. It could therefore be directly related to the concept of mental health. And it may even be compatible with an intelligent, neurophenomenologically optimized form of rational hedonism.

But we may actually be able to go further than this. Obviously, from a more traditional philosophical point of view, the third logical possibility briefly sketched above-minimizing phenomenal transparency-is of greatest interest. Once the principie of autoepistemic closure has been clearly understood on the neurocognitive level, one can define the goal of continuously minimizing the transparency of the PSM. This is in good keeping with the classical philosophical ideal of self-knowledge: To truly accept this ideal means to dissolve any form of autoepistemic closure, on theoretical as well as on phenomenal levels of representation-even if this implies deliberately violating the adaptivity constraint Mother Nature so cruelly imposed on our biological ancestors. Self-knowledge never was a purely theoretical enterprise; it also involves practical neurophenomenologythe sustained effort to epistemically optimize phenomenal self-consciousness itself. It is interesting to note how this traditional principie also unites Eastern and Western philosophy. My prediction is that,

in the centuries to come, the cognitive neuroscience of consciousness will eventually support this old philosophical project of integrating theoretical progress and individual psychological development in a much stronger way than most of us may expect today. The contribution cognitive neuroscience finally makes to the philosophical projects of humanity will be a significant one, because, at its core, cognitive neuroscience is the project of self-knowledge. As I have tried to show in this book, phenomenal selfhood originates in a lack of attentional, subsymbolic self-knowledge. Phenomenal transparency is a special kind of darkness. From a biological point of view this kind of darkness has been enormously successful, because it creates what I have called the "naive-realistic self-misunderstanding." But clearly, from a normative philosophical point of view, representations should always be recognizable as representations and naive realism is something to be abhorred. Eventually, appearance has to be transformed into knowledge.

Perhaps unfortunately, the responsibility of academic philosophy also consists in telling people what they don 't want to hear. Biological evolution is not something to be glorified. It is blind, driven by chance, and it has no mercy. In particular, it is a process that exploits and sacrifices individuals. As soon as individual organisms start to consciously represent themselves as individuals, this fact will inevitably be reflected in countless facets on the level of phenomenal experience itself. Therefore, defining our own goals involves emancipating ourselves from this evolutionary process, which, over millions of years, has shaped the microfunctional landscape of our brains and the representational architecture of our conscious minds. For millions of years, Mother Nature has talked to us, through our reward system and through the emotional layers of our PSM. We have to learn to take a critical stance toward this process, and to view our own phenomenal experience as a direct result of it.' We have to stop glorifying our own neurophenomenological status quo, face the facts, and find the courage lo think about positive alternatives in a rational way. In the end, taking responsibility for the future development of our own conscious minds also is an obvious implication of the project of Enlightenment.

7. Let me give one last example to illustrate the issue: Mother Nature, self-deception, and the emotional selfmodel. The SMT clearly shows how-from a teleofunctionalist perspective-false beliefs about oneself can be functionally adequate relative to a certain environment. Evolution will always have favored those who, until the very last moment, stubbornly believed in themselves. Therefore, the transparency of our PSM may not only be a source of sensory pleasure and self-certainty but a dangerous affair, something frequently depriving us of insight and functional autonomy. To briefly return to an earlier example, the most effective way to deceive others is to deceive yourself as well. In an evolutionary context the causal effect is what counts, not the degree of actual self-knowledge. Consistent self-deception may optimize the genetic success of an organism. Of course, all this will be particularly true of the nonconceptual, for example, the emotional, layers of our self-model as well. Our emotional self-model-one of the central semantic elements in our

traditional folk-psychological notion of the "soul"-may actually be more of a weapon than an instrument to accurately represent reality. It may be something that has arisen from a fundamentally competitive situation, in which cooperation was just one special case of competing in a more intelligent way. In some aspects, having an emotional self-model can even be interpreted as a way of being possessed, possessed by the historical reality that mercilessly burned itself into the inner motivational landscape of our biological ancestors. And the emotional self-model, including all its beautiful aspects, is what drives us. It is a virtual organ ultimately developed to spread genes more efficiently, and not a tool for maximizing self-knowledge. It may therefore make it difficult for us to grasp the true state of affairs, or put already existing insights into action. Any theory about consciousness and the phenomenal self that was maladaptive would immediately be intuitively implausible and emotionally unattractive. For millions of years, Mother Nature has continuously spoken to us through our conscious, emotional self-model. Whenever it is fully transparent, we don't only hear what she says, but we also have the subjective experience of knowing that we know. She says simple things like, "This does not feel right-and you know it doesn't!" or more complicated things like, "Ethical behavior may be the most direct way to make your self-model coherent, yes; but in many situations it will at the sanle time be the most direct way to end your own existence, conscious self and alland you know this is true!" But now we know that the cave is empty. Strictly speaking, there is no one in Che cave who could die. The little red arrow is just a representational device and the pilot is part of the simulator. Strictly speaking, no one was ever boro and no one ever dies. The interesting question is whether purely theoretical points like this one can help us in the situation we now find ourselves in.

Do you recall how, in the first paragraph of the first chapter, I claimed that as you read these lines you constantly confuse yourself with the content of the self-model currently activated by your brain? We now know that this was only an introductory metaphor, because we can now see that this metaphor, if taken too literally, contains a logical mistake: There is no one whose illusion the conscious self could be, no one who is confusing herself with anything. As soon as the basic point has been grasped-the point that the phenomenal self as such is not an epistemically justified form of mental content and that the phenomenal characteristic of selfhood involved results from the transparency of the system model-a new dimension opens. At least in principle, one can wake up from one's biological history. One can grow up, define one's own goals, and become autonomous. And one can start talking back to Mother Nature, elevating her self-conversation to a new level.