1 Questions

1.1 Consciousness, the Phenomenal Self, and the First-Person Perspective

This is a book about consciousness, the phenomenal self, and the first-person perspective. Its main thesis is that no such things as selves exist in the world: Nobody ever *was* or *had* a self. All that ever existed were conscious self-models that could not be recognized *as* models. The phenomenal self is not a thing, but a process—and the subjective experience of *being someone* emerges if a conscious information-processing system operates under a transparent self-model. You are such a system right now, as you read these sentences. Because you cannot recognize your self-model *as* a model, it is transparent: you look right through it. You don't see it. But you see *with* it. In other, more metaphorical, words, the central claim of this book is that as you read these lines you constantly *confuse* yourself with the content of the self-model currently activated by your brain.

This is not your fault. Evolution has made you this way. On the contrary. Arguably, until now, the conscious self-model of human beings is the best invention Mother Nature has made. It is a wonderfully efficient two-way window that allows an organism to conceive of itself *as a whole*, and thereby to causally interact with its inner and outer environment in an entirely new, integrated, and intelligent manner. Consciousness, the phenomenal self, and the first-person perspective are fascinating *representational* phenomena that have a long evolutionary history, a history which eventually led to the formation of complex societies and a cultural embedding of conscious experience itself. For many researchers in the cognitive neurosciences it is now clear that the first-person perspective somehow must have been the decisive link in this transition from biological to cultural evolution. In philosophical quarters, on the other hand, it is popular to say things like "The first-person perspective cannot be reduced to the third-person perspective!" or to develop complex technical arguments showing that some kinds of irreducible first-person facts exist. But nobody ever asks what a first-person perspective *is* in the first place. This is what I will do. I will offer a representationalist and a functionalist analysis of what a consciously experienced first-person perspective is.

This book is also, and in a number of ways, an experiment. You will find conceptual tool kits and new metaphors, case studies of unusual states of mind, as well as multilevel constraints for a comprehensive theory of consciousness. You will find many well-known questions and some preliminary, perhaps even some new answers. On the following pages, I try to build a better bridge—a bridge connecting the humanities and the empirical sciences of the mind more directly. The tool kits and the metaphors, the case studies and the constraints are the very first building blocks for this bridge. What I am interested in is finding conceptually convincing links between subpersonal and personal levels of description, links that at the same time are empirically plausible. What *precisely* is the point at which objective, third-person approaches to the human mind can be integrated with

first-person, subjective, and purely theoretical approaches? How *exactly* does strong, consciously experienced subjectivity emerge out of objective events in the natural world? Today, I believe, this is what we need to know more than anything else.

The epistemic goal of this book consists in finding out whether conscious experience, in particular the experience of *being someone*, resulting from the emergence of a phenomenal self, can be convincingly analyzed on subpersonal levels of description. A related second goal consists in finding out if, and how, our Cartesian intuitions—those deeply entrenched intuitions that tell us that the above-mentioned experience of being a subject and a rational individual can *never* be naturalized or reductively explained—are ultimately rooted in the deeper representational structure of our conscious minds. Intuitions have to be taken seriously. But it is also possible that our best theories about our own minds will turn out to be radically counterintuitive, that they will present us with a new kind of self-knowledge that most of us just cannot believe. Yes, one can certainly look at the current explosion in the mind sciences as a new and breathtaking phase in the pursuit of an old philosophical ideal, the ideal of self-knowledge (see Metzinger, 2000b, p. 6*ff.*). And yes, nobody ever said that a fundamental expansion of knowledge about ourselves necessarily has to be *intuitively* plausible. But if we want it to be a philosophically interesting growth of knowledge, and one that can also be culturally integrated, then we should at least demand an understanding of *why* inevitably it is counterintuitive in some of its aspects. And this problem cannot be solved by any single discipline alone. In order to make progress with regard to the two general epistemic goals just named, we need a better bridge between the humanities and cognitive neuroscience. This is one reason why this book is an experiment, an experiment in *interdisciplinary* philosophy.

In the now flowering interdisciplinary field of research on consciousness there are two rather extreme ways of avoiding the problem. One is the attempt to proceed in a highly pragmatic way, simply generating empirical data without ever getting clear about what the *explanandum* of such an enterprise actually is. The explanandum is that which is to be explained. To give an example, in an important and now classic paper, Francis Crick and Christof Koch introduced the idea of a "neural correlate of consciousness" (Crick and Koch 1990; for further discussion, see Metzinger 2000a). They wrote:

> Everyone has a rough idea of what is meant by consciousness. We feel that it is better to avoid a precise definition of consciousness because of the dangers of premature definition. Until we understand the problem much better, any attempt at a formal definition is likely to be either misleading or overly restrictive, or both. (Crick and Koch 1990, p. 264)

There certainly are a number of good points behind this strategy. In complex domains, as historical experience shows, scientific breakthroughs are frequently achieved simply by stumbling onto highly relevant data, rather than by carrying out rigorously systematized

research programs. Insight often comes as a surprise. From a purely heuristic perspective, narrowing down the scope of one's search too early certainly is dangerous, for instance, by making attempts at excessive, but not yet data-driven formal modeling. A certain degree of open-mindedness is necessary. On the other hand, it is simply not true that everyone has a rough idea of what the term "consciousness" refers to. In my own experience, for example, the most frequent misunderstanding lies in confusing phenomenal experience as such with what philosophers call "reflexive self-consciousness," the actualized capacity to cognitively refer to yourself, using some sort of concept-like or quasi-linguistic kind of mental structure. According to this definition hardly anything on this planet, including many humans during most of their day, is ever conscious at all. Second, in many languages on this planet we do not even find an adequate counterpart for the English term "consciousness" (Wilkes 1988b). Why did all these linguistic communities obviously not see the need for developing a unitary concept of their own? Is it possible that the *phenomenon* did not exist for these communities? And third, it should simply be embarrassing for any scientist to not be able to clearly state *what* it is that she is trying to explain (Bieri 1995). What is the *explanandum*? What are the actual entities between which an explanatory relationship is to be established? Especially when pressed by the humanities, hard scientists should at least be able to state clearly what it is they want to know, what the target of their research is, and what, from their perspective, would count as a successful explanation.

The other extreme is something that is frequently found in philosophy, particularly in the best of philosophy of mind. I call it "analytical scholasticism." It consists in an equally dangerous tendency toward arrogant armchair theorizing, at the same time ignoring first-person phenomenological as well as third-person empirical constraints in the formation of one's basic conceptual tools. In extreme cases, the target domain is treated as if it consisted only of *analysanda*, and not of explananda *and* analysanda. What is an analysandum? An analysandum is a certain way of speaking about a phenomenon, a way that creates logical and intuitive problems. If consciousness and subjectivity were only analysanda, then we could solve all the philosophical puzzles related to consciousness, the phenomenal self, and the first-person perspective by changing the way we talk. We would have to do to modal logic and formal semantics, and not cognitive neuroscience. Philosophy would be a fundamentalist discipline that could decide on the truth and falsity of empirical statements by logical argument alone. I just cannot believe that this should be so.

Certainly by far the best contributions to philosophy of mind in the last century have come from analytical philosophers, philosophers in the tradition of Frege and Wittgenstein. Because many such philosophers are superb at analyzing the deeper structure of language, they often fall into the trap of analyzing the conscious mind as if it were

itself a linguistic entity, based not on dynamical self-organization in the human brain, but on a disembodied system of rule-based information processing. At least they frequently assume that there is a "content level" in the human mind that can be investigated without knowing anything about "vehicle properties," about properties of the actual physical carriers of conscious content. The vehicle-content distinction for mental representations certainly is a powerful tool in many theoretical contexts. But our best and empirically plausible theories of representation, those now so successfully employed in connectionist and dynamicist models of cognitive functioning, show that any philosophical theory of mind treating vehicle and content as anything more than two strongly interrelated aspects of one and the same phenomenon simply deprives itself of much of its explanatory power, if not of its realism and epistemological rationality. The resulting terminologies then are of little relevance to researchers in other fields, as some of their basic assumptions immediately appear ridiculously implausible from an empirical point of view. Because many analytical philosophers are excellent logicians, they also have a tendency to get technical even if there is not yet a point to it—even if there are not yet any data to fill their conceptual structures with content and anchor them in the real-world growth of knowledge. Epistemic progress in the real world is something that is achieved by all disciplines *together*. However, the deeper motive behind falling into the other extreme, the isolationist extreme of sterility and scholasticism, may really be something else. Frequently it may actually be an unacknowledged respect for the rigor, the seriousness, and the true intellectual substance perceived in the hard sciences of the mind. Interestingly, in speaking and listening not only to philosophers but to a number of eminent neuroscientists as well, I have often discovered a "motivational mirror image." As it turns out, many neuroscientists are actually much more philosophers than they would like to admit. The same motivational structure, the same sense of respect exists in empirical investigators avoiding precise definitions: They know too well that deeper methodological and metatheoretical issues exist, and that these issues are important and extremely difficult at the same time. The lesson to be drawn from this situation seems to be simple and clear: somehow the good aspects of both extremes have to be united. And because there already is a deep (if sometimes unadmitted) mutual respect between the disciplines, between the hard sciences of the mind and the humanities, I believe that the chances for building more direct bridges are actually better than some of us think.

As many authors have noted, what is needed is a *middle course* of a yet-to-be-discovered nature. I have tried to steer such a middle course in this book—and I have paid a high price for it, as readers will soon begin to notice. The treatment of philosophical issues will strike all philosophers as much too brief and quite superficial. On the other hand, my selection of empirical constraints, of case studies, and of isolated data points must strike neuro- and cognitive scientists alike as often highly idiosyncratic and quite

badly informed. Yet bridges begin with small stones, and there are only so many stones an individual person can carry. My goal, therefore, is rather modest: If at least *some* of the bits and pieces here assembled are useful to *some* of my readers, then this will be enough.

As everybody knows the problem of consciousness has gained the increasing attention of philosophers (see, e.g., Metzinger 1995a), as well as researchers working in the neuro- and cognitive sciences (see, e.g., Metzinger 2000a), during the last three decades of the twentieth century. We have witnessed a true renaissance. As many have argued, consciousness is the most fascinating research target conceivable, the greatest remaining challenge to the scientific worldview as well as the centerpiece of any philosophical theory of mind. What is it that makes consciousness such a special target phenomenon? In conscious experience *a reality is present*. But what does it mean to say that, for all beings enjoying conscious experience, necessarily *a world appears*? It means at least three different things: In conscious experience there is a world, there is a self, and there is a relation between both—because in an interesting sense this world appears *to* the experiencing self. We can therefore distinguish three different aspects of our original question. The first set of questions is about what it means that a reality *appears*. The second set is about how it can be that this reality appears to *someone*, to a subject of experience. The third set is about how this subject becomes the center of its own world, how it transforms the appearance of a reality into a truly *subjective* phenomenon by tying it to an individual first-person perspective.

I have said a lot about what the problem of consciousness as such amounts to elsewhere (e.g., Metzinger 1995e). The deeper and more specific problem of how one's own personal *identity* appears in conscious experience and how one develops an inward, subjective *perspective* not only toward the external world as such but also to other persons in it and the ongoing internal process of experience itself is what concerns us here. Let us therefore look at the second set of issues. For human beings, during the ongoing process of conscious experience characterizing their waking and dreaming life, *a self is present*. Human beings consciously experience themselves as *being someone*. The conscious experience of being someone, however, has many different aspects—bodily, emotional, and cognitive. In philosophy, as well as in cognitive neuroscience, we have recently witnessed a lot of excellent work focusing on bodily self-experience (see, e.g., Bermúdez, Marcel, and Eilan 1995), on emotional self-consciousness (see, e.g., Damasio 1994, 2000), and on the intricacies involved in cognitive self-reference and the conscious experience of being an embodied *thinking self* (see, e.g., Nagel 1986, Bermúdez 1998). What does it mean to say that, for conscious human beings, *a self is present*? How are the different layers of the embodied, the emotional, and the thinking self connected to each other? How do they influence each other? I prepare some new answers in the second half of this book.

This book, however, is not only about consciousness and self-consciousness. The yet deeper question behind the phenomenal appearance of a world and of a self is connected to the notion of a consciously experienced "first-person perspective": what precisely makes consciousness a *subjective* phenomenon? This is the second half of my first epistemic target. The issue is not only how a phenomenal self per se can arise but how beings like ourselves come to use this phenomenal self as a tool for experiencing themselves as subjects. We need interdisciplinary answers to questions like these: What does it mean that in conscious experience we are not only *related to the world*, but related to it *as knowing selves*? What, exactly, does it mean that a phenomenal self typically is not only present in an experiential reality but that at the same time it forms the *center* of this reality? How do we come to think and speak about ourselves as *first persons*? After first having developed in chapters 2, 3, and 4 some simple tools that help us understand how, more generally, a reality can appear, I proceed to tackle these questions from the second half of chapter 6 onward. More about the architecture of what follows in section 1.3.

1.2 Questions

In this section I want to develop a small and concise set of questions, in order to guide us through the complex theoretical landscape associated with the phenomenon of subjective experience. I promise that in the final chapter of this book I will return to each one of these questions, by giving brief, condensed answers to each of them. The *longer* answers, however, can only be found in the middle chapters of this book. This book is written for readers, and one function of the following minimal catalogue of philosophical problems consists in increasing its usability. However, this small checklist could also function as a starting point for a minimal set of criteria for judging the current status of competing approaches, including the one presented here. How many of these questions can it answer in a satisfactory way? Let us look at them. A first, and basic, group of questions concerns the meaning of some of the explanatory core concepts already introduced above:

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

Alternatively, what does it mean of a conscious system—a person, a biological organism, or an artificial system—if taken as a whole to say that it is conscious?

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

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

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

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

What is a phenomenal first-person perspective, for example, as opposed to a linguistic, cognitive, or epistemic first-person perspective? Is there anything like aperspectival consciousness or even self-consciousness?

Next there is a range of questions concerning ontological, logical-semantic, and epistemological issues. They do not form the focus of this investigation, but they are of great relevance to the bigger picture that could eventually emerge from an empirically based philosophical theory of self-consciousness.

Is the notion of a "subject" logically primitive? Does its existence 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 an entity that could in principle be eliminated in the course of scientific progress?*

In particular, the semantics of the indexical word *I* needs further clarification. What is needed is a better understanding of a certain class of sentences, namely, those in which the word *I* is used in the autophenomenological self-ascription of phenomenal properties (as in "I am feeling a toothache right now").

What are the truth-conditions for sentences of this type?

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

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?

Does the incorrigibility of self-ascriptions of psychological properties imply their infallibility?

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

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?*

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

The true focus of the current proposal, however, is phenomenal content, the way certain representational states *feel* from the first-person perspective. Of particular importance are attempts to shed light on the historical roots of certain philosophical intuitions—like, for

instance, 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 transforming them into more differentiated versions. However, an additional and 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 taking a closer look at the genetic conditions for certain introspective certainties.

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?

How do Cartesian intuitions—like the contingency intuition, the indivisibility intuition, or the intuition of epistemic immediacy—emerge?

Arguably, the human variety of conscious subjectivity is unique on this planet, namely, in that it is culturally embedded, in that it allows not only for introspective but also for linguistic access, and in that the contents of our phenomenal states can also become the target of exclusively internal cognitive self-reference. In particular, it forms the basis of *intersubjective* achievements. The interesting question is how the actual contents of experience *change* through this constant integration into other representational 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?

A final set of phenomenological questions concerns the internal web of relations between certain phenomenal state classes or global phenomenal properties. Here is a brief selection:

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

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?*

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

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

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

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

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

In what 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?*

Finally, one last question concerns the status of *phenomenal universals*: Can we define a notion of consciousness and subjectivity that is hardware- and species-independent? This issue 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, aiming at liberation from any kind of physical domain-specificity. Can there be a *universal* theory of consciousness? In other words:

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

1.3 Overview: The Architecture of the Book

In this book you will find twelve new conceptual instruments, two new theoretical entities, a double set of neurophenomenological case studies, and some heuristic metaphors. Perhaps most important, I introduce two new theoretical entities: the "phenomenal self-model" (PSM; see section 6.1) and the "phenomenal model of the intentionality relation" (PMIR; see section 6.5). I contend that these entities are *distinct* theoretical entities and argue that they may form the decisive conceptual link between first-person and third-person approaches to the conscious mind. I also claim that they are distinct in terms of relating to clearly isolable and correlated phenomena on the phenomenological, the representationalist, the functionalist, and the neurobiological levels of description. A PSM and a PMIR are something to be *found* by empirical research in the mind sciences. Second, these two hypothetical entities are helpful on the level of conceptual analysis as well. They may form the decisive conceptual link between consciousness research in the humanities and consciousness research in the sciences. For philosophy of mind, they serve as important conceptual links between personal and subpersonal levels of description for conscious

systems. Apart from the necessary normative context, what makes a nonperson a person is a very special sort of PSM, plus a PMIR: You *become* a person by possessing a transparent self-model plus a conscious model of the "arrow of intentionality" linking you to the world. In addition, the two new hypothetical entities can further support us in developing an extended representationalist framework for intersubjectivity and social cognition, because they allow us to understand the *second*-person perspective—the consciously experienced *you*—as well. Third, if we want to get a better grasp on the transition from biological to cultural evolution, both entities are likely to constitute important aspects of the actual linkage to be described. And finally, they will also prove to be fruitful in developing a metatheoretical account about what actually it *is* that theories in the neuro- and cognitive sciences are talking about.

As can be seen from what has just been said, chapter 6 is in some ways the most important part of this book, because it explains what a phenomenal self-model and the phenomenal model of the intentionality relation actually are. However, to create some common ground I will start by first introducing some simple tools in the following chapter. In chapter 2 I explain what mental representation is, as opposed to mental *simulation* and mental *presentation*—and what it means that all three phenomena can exist in an unconscious and a conscious form. This chapter is mirrored in chapter 5, which reapplies the new conceptual distinctions to *self*-representation, *self*-simulation, and *self*-presentation. As chapter 2 is of a more introductory character, it also is much longer than chapter 5. Chapter 3 investigates more closely the transition from unconscious information processing in the brain to full-blown phenomenal experience. There, you will find a set of ten constraints, which any mental representation has to satisfy if its content wants to count as *conscious* content. However, as you will discover, some of these constraints are domain-specific, and not all of them form strictly necessary conditions: there are *degrees* of phenomenality. Neither consciousness nor self-consciousness is an all-or-nothing affair. In addition, these constraints are also "multilevel" constraints in that they make an attempt to take the first-person phenomenology, the representational and functional architecture, and the neuroscience of consciousness seriously at the same time. Chapter 3 is mirrored in the first part of chapter 6, namely, in applying these constraints to the special case of self-consciousness. Chapter 4 presents a brief set of neurophenomenological case studies. We take a closer look at interesting clinical phenomena such as agnosia, neglect, blindsight, and hallucinations, and also at ordinary forms of what I call "deviant phenomenal models of reality," for example, dreams. One function of these case studies is to show us what is *not necessary* in the deep structure of conscious experience, and to prevent us from drawing false conclusions on the conceptual level. They also function as a harsh reality test for the philosophical instruments developed in both of the preceding chapters. Chapter 4 is mirrored again in chapter 7. Chapter 7 expands on chapter 4. Because self-

consciousness and the first-person perspective constitute the true thematic focus of this book, our reality test has to be much more extensive in its second half, and harsher too. In particular, we have to see if not only our new set of concepts and constraints but the two central theoretical entities—the PSM and the PMIR, as introduced in chapter 6—actually have a chance to survive any such reality test. Finally, chapter 8 makes an attempt to draw the different threads together in a more general and illustrative manner. It also offers minianswers to the questions listed in the preceding section of this chapter, and some brief concluding remarks about potential future directions.

This book was written for readers, and I have tried to make it as easy to use as possible. Different readers will take different paths. If you have no time to read the entire book, skip to chapter 8 and work your way back where necessary. If you are a philosopher interested in neurophenomenological case studies that challenge traditional theories of the conscious mind, go to chapters 4 and 7. If you are an empirical scientist *or* a philosopher mainly interested in constraints on the notion of conscious representation, go to chapter 3 and then on to sections 6.1 and 6.2 to learn more about the specific application of these constraints in developing a theory of the phenomenal self. If your focus is on the heart of the theory, on the two new theoretical entities called the PSM and the PMIR, then you should simply try to read chapter 6 first. But if you are interested in learning why qualia don't exist, what the actual items in our basic conceptual tool kit are, and why all of this is primarily a *representationalist* theory of consciousness, the phenomenal self, and the first-person perspective, then simply turn this page and go on.