1 Introduction: Consciousness Research at the End of the Twentieth Century

From False Intuitions to Psychophysical Correlations

In 1989 the philosopher Colin McGinn asked the following question: "How can technicolor phenomenology arise from soggy gray matter?" (1989: 349). Since then many authors in the field of consciousness research have quoted this question over and over, like a slogan that in a nutshell conveys a deep and important theoretical problem. It seems that almost none of them discovered the subtle trap inherent in this question. The brain is not gray. The brain is colorless.

Obviously, the fundamental methodological problem faced by any rigorous research program on consciousness is the subjectivity of the target phenomenon. It consists in the simple fact that conscious experience, under standard conditions, is always tied to an individual, first-person perspective. The subjective qualities inherent in a phenomenal color experience are a paradigm example of something that is accessible from a first-person perspective only. Color consciousness—regardless whether in gray or in Technicolor—is a *subjective* phenomenon. However, the precise nature of the relationship of such first-person phenomena to elements within the domain of objectively describable events is unclear. From an objective, third-person perspective all we find in the world are electromagnetic radiation and the reflectance properties of middle-sized objects, wavelength mixtures and metamers, retinal input vectors and activation patterns in the visual system. None of these, so far, map nicely and systematically onto the chromatic primitives of subjective, visual experience. It is just as our physics teacher in high school always told us: From a strictly objective perspective, no such things as colors exist in the world. Therefore, the pivotal question is *not* How do we get from gray to Technicolor?

The core question is if at all—and if so, in what sense—physical states of the human nervous system, under a certain description, can be successfully mapped onto the content of conscious experience. This content can be a simple qualitative feature like "grayness" or "sogginess." There are also complex, nested forms of conscious content like "the self in the act of knowing" (see, e.g., chapters 7 and 20 in this volume) or high-level phenomenal properties like "coherence" or "holism" (e.g., chapters 8 and 9 in this volume). But what, precisely, does it mean that conscious experience has a "content"? Is this an entity open to empirical research programs and interdisciplinary cooperation? And what would it mean to map this content onto physical states "under a certain description"? In other words: What kinds of relations *are* psychophysical correlations? Do we have a workable conception of the isomorphism we are obviously assuming? If one is seriously interested in getting away from the naïveté of popular discussions concerning consciousness, the first thing one has to understand is that we know the world only under representations. For philosophers this is a point of great triviality, but since the large majority of contributors in this volume address empirical issues, a few short remarks may be in order. Let me explain.

Theoretical and Phenomenal Models of Reality

One way to know the world (and ourselves) is under *theoretical* representations. For instance, we can use descriptions of the brain generated by empirical research in the cognitive neurosciences. Neurophysiological descriptions of certain brain areas or neural algorithms describing their computational properties are typical and well-known examples. We can also gain further knowledge under conceptual interpretations of such descriptions generated by analytical philosophers of mind. For instance, philosophers might speak about the way in which a certain abstract property, such as a causal role, is "realized" by a certain concrete state in the brain. Both types of descriptions are linguistic representations, and their content is propositional.

Another way to know the world (and ourselves) is under a *phenomenal* representation. For instance, to come back to our initial example, we can use the content of conscious experience generated by our own brain in the act of visually perceiving another brain in order to gain knowledge about the world. "Grayness," for instance, is one important aspect of the content of a phenomenal representation. The subjectively experienced colors of a rainbow or those of a movie in Technicolor are further examples. The format of phenomenal representations is something for which we currently possess no precise terminology, but it is obviously not of a syntactically structured, linguistic kind, and their content is only very rarely of a conceptual or propositional nature. You don't need language to be conscious—a nonlinguistic creature could certainly have the subjective experience of "grayness."[1] Again, there are also conceptual interpretations of the content of conscious representations itself (for instance, generated by phenomenologically oriented philosophers of mind), and in some cases such descriptions constitute a valuable source of information.

At the end of the twentieth century we have some good ideas about what it could mean for an empirical theory (the first type of representation) to possess "content." However, it is unclear what it means, precisely, to claim that states of consciousness (the second type of representation) have "content." I am not going to answer this question here. But let me frame it in a simplified way that may serve to illustrate an important aspect of the underlying issue. The problem may consist in the fact that phenomenal representations are special in having *two* kinds of content. Philosophers sometimes speak of the *intentional content* and of the *phenomenal content* of mental representations. Consider the following example: While visiting one of the new underground laboratories for experimental philosophy of mind, which are mushrooming all over the world, you suddenly find yourself holding a freshly excised human brain in your hand and, looking at it, you have the phenomenal experience of "grayness" and "sogginess." The next night, after awaking from a nightmare in which you subjectively relived exactly the same scene, including precisely the same visual and tactile qualities, you realize that you have just had a complex hallucination. This time, fortunately, it was all a dream.

What was the difference between the two episodes? In a first and very rough approximation one might say the following: In the initial case your relevant mental state had intentional *and* phenomenal content. The intentional content consisted in the fact that this mental state actually referred to something in the external world; there really *was* a brain in your hand. The phenomenal content consisted, for example, in the subjectively experienced qualities of "grayness" and "sogginess." In the second case, however, there was *only* phenomenal content, because no such thing as a brain existed in your present environment—your hand was paralyzed and your visual system was decoupled from external input (regarding dreams as a model system for phenomenal experience, see chapter 4 in this volume). If you remove the external component, you seem to get very close to the pure experiential content (on the neural correlates of spontaneous visual hallucinations and on bistable phenomena, see chapters 14 and 15 in this volume).

It is probably safe to say that a majority of experts in the relevant areas of philosophy would, while wholeheartedly disagreeing about the nature of intentional content, at least subscribe to the thesis that phenomenal content, in a strong sense, supervenes on properties of the brain.[2] That is, as soon as all internal and contemporaneous properties of your brain are fixed, all properties of your conscious experience are fully determined as well. What is determined is how being in these states *feels* to you, not if these states are what philosophers would call "epistemic states"—states that actually carry knowledge by relating you to the world in a meaningful way. In the short introductions

written for the parts of this volume, I will use the concept of "phenomenal content" in accordance with this loose, nontechnical definition: The phenomenal content of your mental representations is that aspect which, being independent of their veridicality, is available for conscious experience from the first-person perspective while simultaneously being determined by inclusively internal properties of your brain.

What is the upshot of this first conceptual clarification? Consciously experienced colors or the tactile experience of "sogginess" are parts of a *phenomenal* model of reality. The content of global conscious states like waking or dreaming is the content of phenomenal models of reality, episodically activated by the brain of an individual human being. Wavelength mixtures and the like are theoretical entities in *scientific* models of reality. Scientific models of reality are generated by socially interacting groups of human beings. This point is important in order to prevent a second possible form of popular naïveté lurking in the background. The reality of the brain as well as the reality of consciousness as described by science are, strictly speaking, not "the" objective domain. They are the result of *intersubjective* cooperation within scientific communities. If readers will permit the use of a connectionist metaphor: A theoretical model is more like a distributed and coherent pattern in a social network, dynamically unfolding its informational content while subtly changing the internal landscape of the overall system. It is also interesting to note that, in parallel with the renaissance of systematic research programs on conscious experience, we are starting to discover the neural correlates of social cognition as well (see chapter 22 in this volume).

If individual human beings, maybe as observers of a neurosurgical operation or, while in the basement of a pathology institute as witnesses of the dissection of a corpse, consciously look at the exposed brain of a fellow human being, then they will, under standard conditions, experience this brain as having the color gray. *Their* brains activate individual phenomenal models of reality, including the visually perceived brain. From an objective point of view, however, both brains involved in this perceptual relation are absolutely colorless. There are no colors in the external world. Matter never was gray. So what is it that generates those false intuitions often leading us astray? It is the fact that theoretical reality-modeling is anchored in phenomenal reality-modeling, and that phenomenal reality-modeling is characterized by an all-pervading naive realism.

From a strictly subjective point of view there is only one brain, and in all its concrete sogginess and grayness it is certainly not perceived by another brain, but by a self-conscious *person*. This person enjoys what Revonsuo (see chapter 4 in this volume) has called an "out-of-the-brain-experience": a very robust sense of presence in and the immersion into a seemingly real world outside the brain. Yet many theoretical considerations and a flood of empirical data now strongly point to the conclusion that in all its ultimate realism, this form of experiential content is itself entirely dependent on the internal workings of an individual brain. And trying to understand this nexus between the virtuality of an external existence and the internal dynamics of biological information-processing certainly is more exciting than any popular debate could ever be. While the Mysterian's trap is just a rhetorical bogeyman, we are actually faced with much deeper theoretical issues and an extremely interesting set of empirical challenges. This book is about these challenges. How could genuine first-person phenomenal experience emerge in a self-organizing physical universe?

The NCC: Correlating Phenomenal Content with Properties of the Brain

Given this context, what does it mean to look for the "neural correlates of consciousness" (NCC)? The idea of an NCC has been around in dis-

cussions since about 1980, and was probably first used in print by Francis Crick and Christof Koch (1990). In some cases it will mean looking for correlations between certain events in the brain—under a certain representation, as described on a certain neurobiological level of analysis—and for certain events in the ongoing dynamics of phenomenal experience—under a certain representation, as described by the attending, cognizing subject, usually in the everyday terminology of "folk phenomenology." In other cases it will mean looking for correlations between the occurrence of events of the first kind—again, as neuroscientifically described—and the occurrence of events of the second kind—as only *indicated* in a nonlinguistic manner by the subject, such as in pushing a button. Generally speaking, the epistemic goal—what we really want to *know*—in the type of correlation studies relevant to consciousness research consists in isolating the *minimally sufficient neural correlate* for specific kinds of phenomenal content (see chapter 2 in this volume). Such a correlate, however, will always be relative to a certain class of systems and to internal as well as external conditions. In this empirical context it will be the minimal set of properties, described on an appropriate level of neuroscientific analysis, that is sufficient to activate a certain conscious content in the mind of the organism.

However, mapping does not mean reduction. Correlation does not mean explanation. Once strict, fine-grained correlations between brain states and conscious states have been established, a number of theoretical options are still open. Additional constraints therefore will eventually be needed. Important questions are What is the true nature of these psychophysical correlations? Are we justified in interpreting them as *causal* relations? What additional constraints would have to be introduced in order to speak of *lawlike correlations* (see chapter 3 in this volume)? Is a fully reductive account, or even an eliminativist strategy, possible? (See, e.g., P. M. Churchland 1985, 1986, 1988; P. S. Churchland 1986.)

Assume that we find a strict and systematic correlation between a certain brain property or type of neural event N and the subjectively experienced phenomenal property of "sogginess" S. This is entirely compatible with Cartesian dualism: The underlying relation could indeed be a causal one, namely causal *interaction* between events in two ontologically distinct domains. If the ideas of Descartes or those of Popper and Eccles (see Popper and Eccles; Popper 1996) were correct, then we would certainly find neural correlates of consciousness. However, it could also be the case that we have only a unidirectional arrow pointing from N to S, a causal one-way street leading upward from the brain into the conscious mind. If *epiphenomenalism* were true, phenomenal experience as such would be causally inefficacious.[3] Certainly most neuroscientists today would rather be epiphenomenalists than dualists. The problem is this: Empirical correlation data do not help us to decide between those two positions. A third possibility is that there may be no direct causal relationship between N and S at all; they could both be dependent on a single event in the past or upon repeated singular events, constantly reestablishing the observed correlation. The classical position for the first type of interpretation is the Leibnizian concept of prestabilized harmony, the second model is historically represented by "occasionalist" philosophers like Arnold Geulincx and Malebranche, who thought that God would, *ad occasionem*, actively correlate the minds of all human beings with their body whenever necessary. The methodological problem in the background is that of *screening off* N and S from more distant, alternative causes (see chapter 17 in this volume). One typical example of the absence of direct causal relationships between highly correlated sets of events is clocks: Usually large numbers of clocks and watches in our environment all show the same time, though

they do not possess direct causal links in any interesting sense.

If we find strict, fine-grained, and systematic correlations between neural and phenomenal types of events, this does not rule out a fourth possibility. There may be *no* causal relationship between events like N and S at all, neither direct nor indirect, because both of them are just different aspects of one underlying reality. *Double-aspect theories* would assume that scientifically describing N and phenomenally experiencing S are just two different ways of accessing one and the same underlying reality. Spinoza is a beautiful classical example of this philosophical intuition, as is Herbert Feigl with his "neutral monism" version of the early identity theory.[4] Identity theorists frequently thought that the relation between types of mental and physical events was simply that of *contingent identity*. Just as concepts like "morning star" and "evening star" turned out to be coextensive (referring to the same part of reality, the planet Venus), so, they thought, as science advances, mental and physical concepts would in the same way eventually turn out to be referring to one and the same part of reality (see, e.g. Place 1956, 1988). Identity theories are maximally parsimonious and do justice to the principle of the causal closure of the physical world, and as such they seem ideally suited as an underlying research heuristic for the cognitive neurosciences. However, they have their own logical intricacies and difficulties, none of which I am going to discuss here.[5]

What this brief look at some possible ontological interpretations of empirical correlations between mind and brain illustrates is that a full-blown theory of consciousness will need much more than correlation data alone. Taken by themselves, those data simply underdetermine the shape any comprehensive theory will have to take. On the other hand, the work presented in this volume certainly is an impressive beginning. We clearly see a new phase of consciousness research, which is now definitely expanding from the realm of more abstract and speculative models into the field of gathering "harder" and more domain-specific data. And in the end it may even turn out that as we gain new insights about what all those difficult concepts like "first-person perspective," "subjective access," and "introspective individuation of conscious states by their phenomenal content" might actually refer to in terms of their necessary neurocomputational underpinnings, some of our theoretical intuitions about what is really needed to successfully bridge the explanatory gap (see Levine 1983, 1993) will shift as well.

Being a philosopher, I will not attempt to develop a general introduction into what the problem of consciousness amounts to for the cognitive neurosciences.[6] I have given a somewhat more comprehensive introduction into the philosophical issues associated with conscious experience elsewhere (see Metzinger 1995a), and will not repeat myself here.[7] However, let me briefly point to a third possible form of naïveté, which has to be avoided if we want to achieve genuine progress on consciousness.

In order to get away from the shallowness and the constant misunderstandings inherent in many popular discussions of conscious experience, one has to first understand that reduction is a relationship between *theories*, and not between phenomena. A primitive scientistic ideology would be just as bad as succumbing to Mysterianism. Neither serious empirical researcher nor philosopher wants to "reduce consciousness." All that can be reduced is one *theory* about how the contents of conscious experience come into existence to another *theory* about how the contents of conscious experience come into existence. Our theories about the phenomena change. The phenomena stay the same. A beautiful rainbow remains a beautiful rainbow even after an explanation in terms of electromagnetic radiation has become available. Of course, if one takes a second look, it is here where one discovers yet another danger of naïveté lurking in the back-

ground: One factor that makes consciousness such a very special target for scientific research is that our own beliefs about it can subtly *change* the contents and the functional profile of subjective experience itself. Consciousness—as well as science—is a culturally embedded phenomenon (see chapter 8 in this volume).

Soft Issues: The Wider Context of Consciousness Research at the Turn of the Millennium

In the closing section of this general introduction I will briefly draw attention to a number of "soft" issues associated with the search for the NCC. I know that many of my readers will not be interested in these aspects of the problem. They may safely skip the remaining part of this introduction and continue reading in chapter 2. I am also aware that this is a risky enterprise, since there is a rising suspicion about the value of philosophical contributions in consciousness research in general and since a large variety of potential misunderstandings exist. On the other hand, I am convinced that there is an important set of more general and *normative* issues associated with the kind of research now expanding so forcefully. For the twenty-first century's mind sciences, these issues will definitely become more pressing and relevant. They certainly deserve attention. For the scientific community it is vital to keep an eye on these issues from the very beginning, because they will eventually shape our image of ourselves and the cultural foundations of our societies. There is a large normative vacuum emerging, and it is important for it not to be filled by popular irrationalism and by people who are just promoting their own interests and pet ideologies. Those of us who are seriously interested in the growth of knowledge as a good in itself must also face the consequences of this growth. We have to see to it that the ensuing issues in the wider context eventually are resolved with the same degree of professional attention, rationality, and rigorousness which goes into searching for the neural correlates of consciousness. Let me briefly highlight three aspects.

Anthropology Assessment

There is a new image of man emerging, an image that will dramatically contradict almost all traditional images man has made of himself in the course of his cultural history. For instance, to start with a rather trivial point, it will be strictly incompatible with the Christian image of man, as well as with many metaphysical conceptions developed in non-Western religions. Since about 1990 we have learned more about the human brain than in the three preceding centuries. Not only the cognitive neurosciences and consciousness research, but also a growing number of new disciplines like evolutionary psychology, artificial life, and cognitive robotics, are generating a flood of new insights into the foundations of mentality. Implicit in all these new data on the genetic, evolutionary, or neurocomputational roots of conscious human existence is a radically new understanding of what it *means* to be human. Although there is not yet a comprehensive formulation of a new anthropology, the accelerating change in the conception we have of ourselves is becoming more and more obvious. This certainly is an exciting development. As a philosopher, of course, I like to look at it as a new and breathtaking phase in the pursuit of an old philosophical ideal: the ideal of self-knowledge. However, nobody ever said that a deepening of self-knowledge cannot have painful, sobering, or other emotionally unattractive aspects.

Humanity will certainly profit from the current development. But we will also pay a price, and in order to effectively minimize this price, it is important to assess potential consequences of a reductionist neuroanthropology as early as possible. Just as in technology assessment, where one tries to calculate potential dangers, unwanted side-effects and general future consequences of new technologies introduced into society, we

need a new kind of anthropology assessment. We have to start thinking about the consequences a cultural implementation of a new image of man might have.

It may be helpful to differentiate between the "emotional price" and the "sociocultural price." The emotional price consists in a certain unease: We feel insecure, because many of our unscrutinized beliefs about ourselves suddenly seem obsolete. What about rationality and free will—is it really true that our own actions are to a much larger extent determined by "subpersonal" and unconscious events in the brain than we have always liked to assume? If the minimally sufficient neural cause for an overt action and the minimally sufficient neural correlate of the phenomenal experience of *myself now deciding* to carry out this action actually diverge, does this mean that my subjective experience of initiating my own action is some kind of internal confabulation? Is the experience of agency an illusion, a fragile mental construct (see chapter 21 in this volume)? Is conscious thought just a phenomenal echo of the zombie within me talking to itself (see chapter 6 in this volume)? And is there really no such thing as a soul? If the property of selfhood, of "being someone," is not a supernatural essence, but basically a biologically anchored *process* (see chapter 20 in this volume), is there any hope for survival after death? From a purely theoretical perspective the finiteness of human existence in itself does not constitute a problem.

Mortality, however, also is an emotional problem, which we cannot simply brush away by some intellectual operation. The desire for individual survival is one of the highest biological imperatives, mercilessly burned into our limbic system by millions of years of evolution. However, we are the first creature on this planet to have an awareness of the fact that eventually all attempts to observe this bioemotional imperative will be futile. This awareness of mortality will be greatly enhanced as we—especially people outside the academic world and in nondeveloped countries—learn more and more about the neural correlates of consciousness. This is only one element of what I have called the "emotional price." Doubts about the extent to which we actually are free and rational agents are further examples.

There will be a sociocultural price for the current development as well. Unfortunately, this aspect is much harder to assess. First of all, the image we have of ourselves in a subtle, yet very effective, way influences how we live our everyday life and how we interact with our fellow human beings. A popularized form of vulgar materialism following on the heels of neuroscience might therefore lead us into another, reduced kind of social reality. If our image of ourselves is a radically demystified image, then we run the risk of losing a lot of the magic and subtlety in our social relationships. Should believing in a soul or in an irreducible core of our personality one day become just as absurd as stubbornly believing that the sun actually revolves around the Earth is today, then the social and emotional pressures on people who, for whatever reason, have chosen to live their lives outside the scientific image of the world will greatly increase. This may well lead to conflicts, to cultural, and conceivably to civil, warfare. Even today presumably more than 80 percent of the people on this planet do not live their lives against the background of the scientific image of man and, in their personal lives, do not accept even the most general standards of rationality. Almost all of them have never heard of the idea of an NCC, and many of them will not even *want* to hear about it. In short: Existing gaps between the rich, educated, and secularized parts of global society and the poor, less informed, and religiously rooted parts may widen in a way that proves to be unbearable or outright dangerous. One last aspect of the potential sociocultural price to be paid consists in unwanted side effects of new technologies, and they must be rationally assessed and minimized as well.

Consciousness Ethics

We are currently witnessing the beginning of a truly revolutionary development: Subjective experience becomes technologically accessible, in a way it has never been in the history of mankind. This is particularly obvious in the thematic context of this book. Once we know the neural correlate of a specific kind of phenomenal content, we can, in principle, selectively switch this content on and off (see chapters 16–19 in this volume). We can start to modulate it, amplify it, and arguably we can even *multiply* it in artificial systems by realizing the same computational function, the same causal role on another kind of physical hardware. Biological psychiatry, neuropharmacology, and medical neurotechnology, as today manifested in new forms of short-term psychotherapy or new generations of mood enhancers, in the transplantation of embryonic nerve cell tissue or the implantation of brain prostheses, are just the tip of the iceberg. Many of the neuro- and information-processing technologies of the future are going to be *consciousness technologies*, because their main goal will be to directly change the phenomenal content of their targets' mental states. In psychiatry and other branches of medicine this will certainly be a blessing for generations of patients to come. But as it becomes possible to influence and manipulate conscious experience in ever more precise and reliable ways, we face a new ethical dimension. Therefore, more than a research ethics for the cognitive neurosciences or an applied ethics for neurotechnology is needed. We may have to go beyond the concept of mental health used in medicine or psychiatry, and start thinking about what states of consciousness are interesting or desirable *in principle*.

Developing a normative theory of conscious states would be a difficult problem in many respects. First, it would mean constructing a theory that offers not normative judgements of actions, but a normative evaluation of *ways of subjectively experiencing the world*. Maybe one could analyze consciousness ethics as a new branch of ethics dealing with actions having the primary goal of deliberately changing the phenomenal content of mental states possessed by the agent or other persons. Of course, many people have long been seeking a convincing theory about what good and desirable states of consciousness actually are. But it is far from clear if searching for such a theory is even a coherent goal. Does it really make sense to speak of a "good" state of consciousness? In everyday life, are there really states of subjective experience that are "better" than others? A general ethics for conscious experience would inevitably have to face all the foundational issues concerning the epistemic status and the universalizability of ethical norms, which any moral philosophy has to confront. Personally, I tend to be rather skeptical with regard to the prospects of such an ethics for consciousness.

However, decisions will have to be made. And it is interesting to note how large the scope of normative considerations in this realm would be. They would range from pedagogics to euthanasia, from animal rights to robotics, and from drug policy to media policy. It is also surprising to see how far concrete questions range; an ethics of consciousness could attempt to answer them in a more systematic way: What states of consciousness do we want to show our children? What state of consciousness do we eventually want to die in? What states of consciousness would we like to be illegal in our societies? What types of conscious experience do we want to foster and integrate into our culture? What states of consciousness are we allowed to force on animals (e.g., when attempting to isolate the NCC)? Should we really try to build conscious machines before we have understood why our own form of subjective experience is accompanied by so much suffering? How can we design media environments so that they do not endanger our mental health, but increase our own autonomy and the quality of our conscious lives? If we have answers to these questions, we may soon

be able to achieve practical solutions in a more efficient way—by bringing about the NCC of the desired phenomenal state. We might then move on by seeking answers to questions of a more pragmatic kind: How can scientific research on consciousness help us to realize our normative goals? How can we use this research to further minimize the price we pay as much as possible?

Consciousness Culture

Anthropology assessment and ethical considerations are not enough. The issue is not just how to avoid the adverse side effects of a very special and controversial kind of scientific progress. Rather, the crucial point is that new insights about the structure of mind and the wealth of knowledge generated by empirical research on the phenomenon of conscious experience *themselves* have to be culturally implemented. We have to move away from a purely defensive position (as is currently widespread in the humanities), away from any cheap, counterproductive resentment. Laying the foundations for a consciousness culture means taking a more active attitude, a—nevertheless critical—point of view that allows us to ask positive questions like How would a future culture look that uses the results of consciousness research in a fruitful way? Can a *positive* vision be developed? How to protect the individual from new potentials for manipulation and the dangerous side effects of commercially exploited, newly emerging consciousness technologies is just one half of the challenge we will be facing in the future. The other half consists in using those new insights and technologies to *raise* the degree of individual autonomy, in order to help individual human beings live in the states of consciousness in which they have decided to live. Obviously, one necessary precondition consists in being ready to face the facts. *Ought* implies *can*, and objective knowledge is important for any realistic judgement of the options open to us.

A consciousness culture will have nothing to do with organized religion or a specific political vision. Rather, it has to be a rational and productive strategy to transfer new knowledge and new possibilities for action into a global sociocultural context. New knowledge and new technologies, which doubtless, with ever-accelerating speed, will emerge from research activities in the empirical mind-sciences in the next millennium, have to be integrated into society in a way that gives a maximum of people free access to them. A rational consciousness culture, it seems safe to say, will always have to encourage individuals to take responsibility for their own lives—and make continuous attempts at creating a social context that allows them to actually do so. Our current lack of a genuine consciousness culture can be interpreted as an expression of the fact that the project of enlightenment got stuck. What we need is not faith, but knowledge; what we are lacking is not a new metaphysics, but a new variant of practical rationality. In short, the third bundle of ''soft issues'' to which I briefly wanted to point at the end of this introduction is constituted by the urgent necessity to *embed* the current technological and the current theoretical development in a sustainable process of cultural evolution that can keep pace with stormy future developments. It has not been my intention to make any positive suggestions here. All I want to do is throw some light on the broader context in which the search for the NCC is taking place at the turn of the millennium.

However, consciousness culture, just like self-knowledge, is an old philosophical project. Cicero (1971; *Tusculanae disputationes*, II 5) conceived of philosophy as *cultura animi*, as taking care of and cultivating the soul—and in this sense I have only advertized a very old concept of philosophy that went out of fashion a long time ago. Maybe defining the love of wisdom as cultivating the soul is a classical motif that could inspire us as we take our first steps in the present situation. One has to admit, though, that the initial conditions for the time-honored

project of a consciousness culture have changed slightly since the time of Cicero. It therefore remains an open question whether a convincing new interpretation of this classical motif, in light of our recent discoveries about the neurobiological foundations of consciousness and subjective experience, could actually be achieved.

Notes

1. The important question, which I am deliberately skipping in this short introduction, runs in the opposite direction: Could we coherently conceive of a class of representational systems that *only* knows the world under theoretical propositional representations, never having had any kind of subjective experience? In other words, Could the epistemic projects of science and philosophy, at least in principle, be successfully pursued by an unconscious race of machines? Or are even the meaning and the truth of scientific theories ultimately constituted by the fact that they are generated in groups of *phenomenal subjects*—systems that also know the world (and themselves) under phenomenal representations?

2. In philosophy of mind, the concept of supervenience stands for an attempt to formulate a coherent and *nonreductive* form of materialism, capturing the essential theoretical intuitions behind many previous strategies for solving the mind–body problem. For the concept of supervenience, see Kim 1993. For an excellent and accessible introduction to philosophy of mind, well suited for empirical researchers and other nonphilosophers, see Kim 1996.

3. Here the classical position is Thomas Huxley's. For a recent exposition of problems surrounding the notion of epiphenomenalism, see Bieri 1992. Herbert Feigl saw the problem of introducing "nomological danglers," a new class of psychophysical laws "dangling out of" the closed causal network of the physical world, as early as 1960: "These correspondence laws are peculiar in that they may be said to postulate 'effects' (mental states as dependent variables) which by themselves do not function, or at least do not seem to be needed, as 'causes' (independent variables) for any observable behaviour" (Feigl 1960: 37).

4. See Feigl 1958; for a collection of texts regarding early identity theory, see Borst 1970.

5. Regarding formal and semantic difficulties of the identity theory, see Kripke 1971, 1972; for the more influential "multiple realization argument" see Putnam 1975, 1992; for a brief introduction to functionalism Block 1980. A good way to enter the current debate is Kim 1998. Important edited collections are Borst 1970; Heil and Mele 1993; Lycan 1990; Warner and Szubka 1994.

6. See Bock and Marsh 1993; Cohen and Schooler 1997; Davies and Humphreys 1993; Marcel and Bisiach 1988; Milner and Rugg 1992 for edited collections. Examples of important individual contributions are Shallice 1988; Weiskrantz 1997; see also the references to monographs given in the introductions to individual parts of this book.

7. An excellent, recent introduction is Güzeldere 1997. For substantial encyclopedia articles, containing further references, see Diemer 1971; Grauman 1966; Landesman 1967; Lormand 1998; Metzinger and Schumacher 1999; NN 1904.

References

Bieri, P. (1992). Trying out epiphenomenalism. *Erkenntnis* 36: 283–309.

Block, N. (1980). What is functionalism? In Block, N., ed. (1980). *Readings in the Philosophy of Psychology. Vol. 1.* Cambridge, MA: Harvard University Press.

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

Bock, G. R., and Marsh, J., eds. (1993). *Experimental and Theoretical Studies of Consciousness*. New York: Wiley.

Borst, C. V., ed. (1970). *The Mind|Brain Identity Theory*. London: Macmillan.

Churchland, P. M. (1985). Reduction, qualia, and the direct introspection of brain states. *Journal of Philosophy* 82: 8–28.

Churchland, P. M. (1986). Some reductive strategies in cognitive neurobiology. *Mind* 95: 279–309. Reprinted in *A Neurocomputational Perspective*. Cambridge, MA: MIT Press.

Churchland, P. M. (1988). Reduction and the neurobiological basis of consciousness. In A. Marcel and E. Bisiach, eds., *Consciousness in Contemporary Science*. Oxford: Oxford University Press.

Churchland, P. S. (1986). *Neurophilosophy: Toward a Unified Science of the Mind–Brain*. Cambridge, MA: MIT Press.

Cicero, Marcus Tullius (1971). *Tusculan disputations*. Loeb classical library; Cambridge, Mass.: Harvard University Press.

Cohen, J. D. and Schooler, J. W., eds. (1997). *Scientific Approaches to Consciousness*. Mahwah, NJ: Lawrence Erlbaum Associates.

Crick, F. H. C., and Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences* 2: 263–275.

Davies, M., and Humphreys, G., eds. (1993). *Consciousness: Psychological and Philosophical Essays*. Oxford: Basil Blackwell.

Diemer, A. (1971). Bewußtsein. In J. Ritter, ed., *Historisches Wörterbuch der Philosophie*. Vol. 1. Basel: Schwabe Verlag.

Feigl, H. (1958). *The "Mental" and the "Physical."* In H. Feigl, M. Scriven and G. Maxwell, eds., *Minnesota Studies in the Philosophy of Science: Concepts, Theories and the Mind-Body-Problem*, Vol. 2. Minneapolis: University of Minneapolis Press.

Feigl, H. (1960). Mind–body, *not* a Pseudo-Problem. In S. Hook, ed., *Dimensions of Mind*. New York: Collier Macmillan.

Graumann, C.-F. (1966). Bewußtsein und Bewußtheit. Probleme und Befunde der psychologischen Bewußtseinsforschung. In W. Metzger and H. Erke, eds., *Allgemeine Psychologie: Vol. I: Der Aufbau des Erkennens*. vol. 1 of K. Gottschaldt et al., eds., *Handbuch der Psychologie*. Göttingen: Verlag für Psychologie.

Güzeldere, G. (1997). Introduction: The many faces of consciousness: A field guide. In Block et al. 1997.

Heil, J., and Mele, A., eds. (1993). *Mental Causation*. Oxford: Clarendon Press.

Kim, J. (1993). *Supervenience and Mind*. Cambridge: Cambridge University Press.

Kim, J. (1996). *Philosophy of Mind*. Boulder, CO: Westview Press.

Kim, J. (1998). *Mind in a Physical World. An Essay on the Mind–Body Problem and Mental Causation*. Cambridge, MA: MIT Press.

Kripke, S. (1971). Identity and necessity. In M. Munitz, ed., *Identity and Individuation*. New York: New York University Press.

Kripke, S. (1972). Naming and necessity. In D. Davidson and G. Harman, eds., *Semantics of Natural Language*. Dordrecht: Reidel Publishing Company. Revised version as monograph (1980), *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Landesman, C., Jr. (1967). Consciousness. In P. Edwards, ed., *The Encyclopedia of Philosophy*. Vol. 2. New York: Macmillan/Free Press.

Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, 64: 354–61.

Levine, J. (1993). On leaving out what it's like. In Davies and Humphreys 1993. Reprinted in Block et al. 1997.

Lormand, E. (1998). Consciousness. In E. Craig and L. Floridi, eds., *Routledge Encyclopedia of Philosophy*. London: Routledge.

Lycan, W. G., ed. (1990). *Mind and Cognition*. Oxford: Basil Blackwell.

Marcel, A., and Bisiach, E., eds. (1988). *Consciousness in Contemporary Science*. Oxford: Oxford University Press.

McGinn, C. (1989). Can we solve the mind–body problem? *Mind* 98: 349–366. Reprinted in Block et al. 1997.

Metzinger, T. (1995a). Introduction: The problem of consciousness. In Metzinger 1995b.

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

Metzinger, T., and Schumacher, R. (1999). Bewußtsein. In H.-J. Sandkühler, ed., *Enzyklopädie der Philosophie*. Hamburg: Meiner.

Milner, D., and Rugg, M., eds. (1992). *The Neuropsychology of Consciousness*. London: Academic Press.

NN. (1904). Bewußtsein. In R. Eisler, ed., *Wörterbuch der philosophischen Begriffe*. Berlin: Ernst Siegfried Mittler und Sohn.

Place, U. T. (1956). Is consciousness a brain process? *British Journal of Psychology* 47: 44–50. Reprinted in Borst 1970.

Place, U. T. (1988). Thirty years on—Is consciousness still a brain process? *Australasian Journal of Philosophy* 66: 208–219.

Popper, K. R. (1996). *Knowledge and the Body–Mind Problem: In Defence of Interaction*. London: Routledge.

Popper, K. R., and Eccles, J. C. (1977). *The Self and Its Brain: An Argument for Interactionism*. Berlin, Heidelberg, London, New York: Springer.

Putnam, H. (1975). *Mind, Language, and Reality*. Vol. 2 of his *Philosophical Papers*. Cambridge, UK: Cambridge University Press.

Putnam, H. (1992). Why functionalism didn't work. In J. Earman, ed., *Inference, Explanation and Other Frustrations. Essays in the Philosophy of Science*. Berkeley: University of California Press.

Shallice, T. (1988). *From Neuropsychology to Mental Structure*. Cambridge: Cambridge University Press.

Warner, R., and Szubka, T., eds. (1994). *The Mind–Body Problem. A Guide to the Current Debate*. Oxford: Basil Blackwell.

Weiskrantz, L. (1997). *Consciousness Lost and Found: A Neuropsychological Exploration*. Oxford: Oxford University Press.

I FOUNDATIONAL ISSUES AND CONCEPTUAL PROBLEMS

David Chalmers and Ansgar Beckermann, the first two authors in this introductory part of the book, are philosophers. Antti Revonsuo and Gerhard Roth, the two contributors following them, are philosophers who in a later phase of their research career became neuroscientists. Their contributions will guide readers into the three middle sections of this volume. Almost all chapters in this middle part focus on the empirical aspects of the ongoing search for the neural correlates of consciousness. However, as readers will undoubtedly notice, many authors turn out to be deeply sensitive to the more theoretical and metatheoretical issues associated with this newly emerging field of research. The final section of this collection will round off the debate by returning to questions of a philosophical and more speculative kind.

What do we actually mean by the concept of a "neural correlate of consciousness"? David Chalmers offers an introductory analysis of this concept and of some of the more general philosophical and methodological issues associated with it. If a neural correlate of consciousness is taken as a specific system in the brain whose activity directly correlates with states of conscious experience, then two questions immediately follow: What is a "state of consciousness"? And what makes a correlation a "direct correlation"?

Chalmers points out that we will often be interested in the correlates of specific types of *phenomenal content* (see chapter 1). The crucial question, as he puts it, is whether the representational content in the neural system matches up with the representational content in, for instance, visual consciousness. States of consciousness, in this way of thinking about them, are individuated by their experiential content, by *what* is subjectively experienced through them. Does "direct correlation" mean that we are looking for neural systems that are necessary and sufficient for consciousness? Chalmers thinks this is too strong a requirement, since it might turn out that there exists more than one neural correlate of a given conscious state. There might, for example, be two systems M and N such that a certain state of M suffices for being in pain and a certain state of N also suffices for being in pain. If we would want to say that both M and N count as neural correlates of this specific conscious content, then both of them would be sufficient but neither would be necessary.

The interesting concept, however, is not merely that of a sufficient neural correlate of consciousness. We do not want irrelevant brain properties to enter into our description of this correlate. What we should be looking for is a *minimally sufficient neural system*. It is defined by (a) being sufficient to bring about the corresponding state of consciousness and (b) the fact that no proper *part* of it suffices by itself to bring about this corresponding state of consciousness. After this important conceptual tool has been established, Chalmers goes on to investigate the domain, the relevant range of cases and conditions under which such a tool can be applied. He closes by offering a series of methodological outcomes from a philosophical perspective.

Once correlations between neural and phenomenal states have been achieved, we face another deep theoretical problem: the explanatory gap (see Levine 1983, 1993). Since nothing in the physical or functional correlates of a phenomenal state helps us to understand why this state subjectively *feels* in a certain way, a special sort of "intelligibility gap" arises. Why so? Phenomenal states are not fully characterized by the causal role the play (e.g., in the generation of behavior). They also have a distinct qualitative character, and many of us can always imagine that whatever realizes the causal role in the brain can be separated from this qualitative, subjective content. There seems to be no *necessary* connection (e.g., from a certain activation pattern in the visual system to *this* specific shade of indigo I am experiencing now). This intuitive separability is one major root of Cartesian intuitions in the philosophy of mind: Reductive strategies to explain qualia and consciousness seem to leave a gap in the explanation, in that, strictly speaking,

such explanations cannot really be *understood*. They do not seem to us to say, in principle, what we want to know. In order to overcome this difficulty, we need a much deeper understanding of the logic behind psychophysical laws; we need an understanding of what it would mean to possess general bridge principles connecting brain states to states of conscious experience.

Ansgar Beckermann in his contribution offers a careful analysis showing how the current theoretical debate is deeply rooted in discussions about the concept of "emergence," which took place at the beginning of the twentieth century.

Could a phenomenal quality—like the one given in the visual experience of a certain shade of indigo—be an *emergent* property in the sense that (a) it is a true law of nature that all brains with a certain microstructure will generate the conscious experience of indigo, while (b) the occurrence of an indigo-experience cannot (not even in principle) be deduced from the most complete knowledge of the properties possessed by all the neural components making up the microstructure, either in isolation or within other arrangements? This was C. D. Broad's definition in his famous book *The Mind and Its Place in Nature*. What are the laws connecting properties of parts to properties of complex wholes? Are they, in the case of phenomenal experience, unique and ultimate laws, which cannot be derived from the general laws of nature? Could a Martian consciousness researcher, who had complete scientific knowledge about the brains of humans but no visual modality, maybe not a even a nervous system, *predict* the occurrence of a sensation of indigo? Or would she be unable even to form a concept of the sensation of indigo before having experienced it at least once? Beckermann offers a clear exposition of the problems we currently face when trying to take the step from empirical correlation studies to fully reductive and genuinely explanatory accounts of phenomenal experience. However, he also remarks that the fact of the non-deducibility of the qualitative character possessed by conscious states from the current laws of neurobiology may simply be a *historical* fact—reflecting an insufficiently advanced state of the neuroscience of consciousness.

This is the point of departure for Antti Revonsuo. He asks what it would take to finally transform the current state of consciousness studies into a rigorous scientific research program on consciousness. We are presently witnessing what in theory of science would be called a "preparadigmatic stage": There is no one coherent theory of consciousness that could serve as the unified background for criticism and systematically organized further developments. Revonsuo points out that what we should strive for is first and foremost a *biological* theory of consciousness, an empirically based strategy that regards consciousness primarily as a biological phenomenon. However, although the volume of empirical research relevant for understanding phenomenal experience in cognitive neuroscience is so large that it is probably the best starting place for the enterprise in general, the metaphysics of cognitive neuroscience appears to be merely some vague and philosophically outdated version of functionalism.

Revonsuo proceeds to isolate some of the major problems that have to be solved. First of all, the initial assumptions of a scientific research program for consciousness have to be clearly formulated. The ontological assumptions for a theory of consciousness have to be conceptually framed in a manner acceptable for a biological research program. Since the corresponding assumptions in cognitive neuroscience are inappropriate, novel levels of description and explanation have to be introduced. At least one level has to capture the first-person point of view (see chapter 20, this volume). Also, a resolution of the "paradox of isomorphism" is needed: If consciousness, as Revonsuo assumes, resides strictly in the brain, then there must be some level of organization in the brain that quite directly resembles the content of conscious experience. In developing initial answers to these

questions, he proposes that, in accordance with what he terms the "standard hierarchical model of the world", a serious research program should reconceptualize consciousness as the phenomenal level of organization in the brain. A complete description of the brain, Revonsuo argues, *necessarily includes* this level of description, which cannot be imported from any other discipline but must be contributed by the science of consciousness itself. In fleshing out his own proposals he then investigates the role of dreams as a model system for consciousness and as a metaphor for subjective experience in general. The conscious brain, in this view, is nature's own virtual reality system that creates an "out-of-the-brain-experience" for the organism, so that the individual can act in a meaningful way in the world. The main task for the research program on consciousness is to describe the phenomenal level systematically, and to capture it through empirical investigations.

The last contribution in this introductory section leads the reader from the philosophy of consciousness into the realm of empirical research. Gerhard Roth draws our attention to the *historical dimension* associated with conscious experience. The human brain can be seen as the result of many millions of years of biological evolution, and it is rational to assume that this is also true of at least major portions of the neural correlates of consciousness embedded in this brain. In short, ontogenetically as well as phylogenetically speaking, conscious experience is an acquired phenomenon. There will be *stages* in which it has naturally developed. If we look across species boundaries and into the evolutionary history of nervous systems on this planet, we will very likely find simple and complex, older and more recent, general and strictly species-specific NCCs as well as types of phenomenal content going along with them. But how are we—foundational issues in philosophy put aside for now—going to find an answer to the question of whether members of other biological species have phenomenal experiences as well? Roth proposes a number of strategies: (1) to check groups of animals for the presence of those cognitive functions which in humans can be exerted only consciously; (2) to examine which parts of the human brain are necessary for (and active during) the different states of consciousness; (3) to examine which of these centers of the human brain are present (and active) in the brains of those animals which—based on behavioral evidence—show certain states of consciousness; (4) to compare the ontogeny of cognitive functions, including states of consciousness in humans, with the ontogeny of the human brain. In the ideal case, the first appearance of certain states of human consciousness should coincide with the maturation of certain centers in the human brain.

If we look at the project of correlating conscious experience with its physical substrates from this new angle, it becomes obvious that any fine-grained analysis will demonstrate that there are not only many different stages in the development of a hypothetical NCC, but also a wider variety of phenomenal states than many of us may have previously thought. The complexity of the research domain called "conscious experience" is rooted in the complexity of its history and in the structural richness of forms brought about by the evolution of life on our planet.

Further Reading

Beckermann, A. (1992). Supervenience, emergence, and reduction. In A. Beckermann, H. Flohr, and J. Kim, eds., *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism*. Berlin: Walter de Gruyter.

Beckermann, A. (1997). Property physicalism, reduction and realization. In M. Carrier and P. Machamer, eds., *Mindscapes: Philosophy, Science, and the Mind*. Konstanz: Universitätsverlag; Pittsburgh: University of Pittsburgh Press.

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

Broad, C. D. (1925). *The Mind and Its Place In Nature*. London: Kegan Paul, Trench, Turbner, and Co.

Byrne, R. (1996). *The Thinking Ape: Evolutionary Origins of Intelligence*. Oxford: Oxford University Press.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness* Studies 2: 200–219. Reprinted in S. Hameroff, A. Kaszniak, and A. Scott, eds. (1996). *Toward a Science of Consciousness*, (Cambridge, MA: MIT Press); and in J. Shear, ed. (1997). *Explaining Consciousness: The Hard Problem*. Cambridge, MA: MIT Press.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Chalmers, D. J. (1998). On the search for the neural correlate of consciousness. In S. Hameroff, A. Kaszniak, and A. Scott, eds., *Toward a Science of Consciousness II*. Cambridge, MA: MIT Press.

Davies, M., and Humphreys, G., eds. (1993). *Consciousness: Psychological and Philosophical Essays*. Oxford: Basil Blackwell.

Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly* 64: 354–361.

Levine, J. (1993). On leaving out what it's like. In Davies and Humphreys 1993. Reprinted in Block et al. 1997.

Pearce, J. M. (1997). *Animal Learning and Cognition*. 2nd ed. Hove, UK: Psychology Press.

Revonsuo, A. (1994). In search of the science of consciousness. In A. Revonsuo and M. Kamppinen, eds., *Consciousness in Philosophy and Cognitive Neuroscience*. Hillsdale, NJ: Lawrence Erlbaum.

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

Revonsuo, A., Wilenius-Emet, M., Kuusela, J., and Lehto, M. (1997). The neural generation of a unified illusion in human vision. *NeuroReport* 8: 3867–3870.

Roth, G. (1998). *Das Gehirn und seine Wirklichkeit: Kognitive Neurobiologie und ihre philosophischen Konsequenzen*. Frankfurt am Main: Suhrkamp.

Roth, G., and Prinz, W., eds. (1996). *Kopf-Arbeit: Gehirnfunktionen und kognitive Leistungen*. Heidelberg: Spektrum Verlag.