Original Paper UDC 165.62: 165.12
Received October 16th, 2007

Bruce J. MacLennan

Department of Electrical Engineering & Computer Science, University of Tennessee, 1123 Volunteer Blvd.
USA-Knoxville, TN 37996–3450
maclennan@eecs.utk.edu

Consciousness: Natural and Artificial

Abstract

Based on results from evolutionary psychology, we discuss important functions that can be served by consciousness in autonomous robots. These include deliberately controlled action, conscious awareness, self-awareness, metacognition, and ego consciousness. We distinguish intrinsic intentionality from consciousness, but argue it is also important to understanding robot cognition. Finally, we explore the Hard Problem for robots (i.e., whether they can experience subjective awareness) from the perspective of the theory of protophenomena.

Keywords

autonomous robot, awareness, consciousness, evolutionary psychology, the Hard Problem, intentionality, metacognition, protophenomena, qualia, synthetic ethology

1. Introduction

There are many scientific and philosophical problems concerning consciousness, but in 1995 David Chalmers proposed using "the Hard Problem" to refer to the principal scientific problem of consciousness, which is to understand how physical processes in the brain relate to subjective experience, to the feeling of being someone. As he put it, "It is widely agreed that experience arises from a physical basis, but we have no good explanation of why and how it so arises" (1995). Unfortunately, the scientific investigation of experience is impeded by the unique epistemological status of consciousness, for while scientific observation is based on specific experiences, consciousness is the ground of all possible experience (MacLennan, 1995). Chalmers called on researchers to face up to the Hard Problem, and Shear (1997) collects a number of papers responding to his challenge.

Of course, neither Chalmers nor I intend to suggest that all the other problems connected with consciousness are "easy"; indeed, some of them are as difficult as any in neuropsychology. However, they may be approached using ordinary scientific methodology, as developed in cognitive science and neuroscience, and so in this sense they are "less hard" than the Hard Problem. They have in common that, at least in principle, they can be solved in terms of neural information processing and control, without reference to any associated subjective experience. In this article I will begin by considering some of these "less hard" problems, especially in the context of robot consciousness, which provides a useful point of contrast and comparison to these problems in the context of humans and other animals. Then I turn to the Hard Problem in the contexts of both natural and artificial intelligence.

2. Less Hard Problems

Functional Consciousness

One of the difficulties in the scientific study of consciousness is that even psychologists and philosophers use the term with a variety of interrelated and overlapping meanings (e.g., Block, 1995). In this section I will consider several of these notions and the "less hard" problems associated with them in the context of animals and robots.

What is consciousness good for? Why has it evolved? Is there any reason we should want robots to be conscious? To answer these questions, we need to understand the *function*, that is, the *purpose fulfilled*, by biological consciousness. In biology, questions of the function of an organ or process are answered by investigating its adaptive value, that is, by asking what selective advantage it has conferred in the species' *environment of evolutionary adaptedness* (EEA), which is the environment in which the species evolved and to which it is adapted. To this end, comparative studies between species are often informative. *Evolutionary psychology* refers to the application of evolutionary biology to psychological questions, and I will use a similar approach to address the "less hard" problems of robot consciousness.[1]

One of the functions of consciousness is to control what is referred to as *voluntary action*, but to avoid irrelevant issues of "free will", it is perhaps less confusing to call it *deliberately controlled action*. Much of our everyday activity is *automatically controlled*, that is, the detailed sensorimotor control is unconscious. Examples include walking, feeding and washing ourselves, and driving a car under ordinary conditions. Under some conditions, however, our control of our actions becomes very conscious and deliberate. This may be required when conditions are abnormal (e.g., walking when you are dizzy or crossing ice, driving in bad weather or traffic), or when we are learning a new skill (which, therefore, is not yet automatic). For example, an unexpected sensation during automatic behavior can trigger an orienting response and a breakdown in the automatized behavior so that it may be placed under more deliberate ("voluntary") control. For example, when walking a leg gets caught or stuck, or the ground is infirm. This may trigger deliberate activity to free the leg or to inspect the local environment. Under *breakdown* conditions we pay much more attention, investing scarce cognitive resources in careful coordination of sensory input and motor behavior; we cannot depend on learned automatic behaviors, with their limited assessments of relevance and programmatic control of response, to do the right thing.

In the terms of Heidegger's phenomenology (e.g., Heidegger, 1962, 1982; Dreyfus, 1991, ch. 4), in our ordinary mode of skilful coping, we encounter objects and the world as *ready-to-hand* (*zuhanden*), in effect as equipment through which our "projects" are being realized. However, when there is a *break* (*Bruch*), that is, a *breakdown* or *disturbance* in this transparent behavior, then the absent or obstructing object or condition becomes *unready-to-hand* (*unzuhanden*) and as a consequence *conspicuous* (*auffällig*). There is a shift from *absorbed coping* to *deliberate coping*. The obstructing object or condition is then encountered as *present-at-hand* (*vorhanden*), in effect, as a problem to be solved so there may be a return to the project. It is treated as a thing or objective situation rather than as equipment or a means appropriate to an end.

Similar considerations apply to autonomous robots when they are operating under exceptional circumstances or learning new skills, and so they should be

able to exert deliberate control over activities that are otherwise automatic, or that may be so once learned. Deliberate control involves the integration of a wider range of information than automatic control (for the latter focuses on information whose relevance has been established) and involves the use of feedback from a wider variety of sources to control action. Information representation is less specific, more general-purpose (and therefore more expensive in terms of neural processing resources). Automatic action makes use of more narrowly focused information representations and processing pathways.

One of the ways that consciousness can facilitate deliberately controlled action is by a process termed *conscious awareness*, that is, by integrating information from memory and various sensory modalities (e.g., visual and kinesthetic), and by using it for more detailed, explicit motor control, which is another function of consciousness. Normally we want automatically controlled activities to take place in more peripheral processing systems involving only the information resources required for their skillful execution, thus leaving the centralized resources of conscious awareness available for higher level processes.

Human beings, and probably many other species, exhibit *visual dominance*, that is, information integration is accomplished by relating it to visual representations. Thus, sounds, odors, tactile perceptions, etc. are bound to parts of visual perceptions and localized with respect to visually perceived space. Memory may trigger these bindings (e.g., the appearance of a hostile agent to its sound) on the basis of stored associations. The fundamental reason for *visual* dominance (as opposed to some other sensory modality) can be found in the shortness of optical wavelengths, which permits detailed imaging of remote objects. The same considerations apply to robots, which suggests that visual dominance may be a good basis for information integration in artificial conscious awareness (in the sense defined above).

Another function of consciousness is *self-awareness*, which in this context does not refer to the ability to contemplate the existential dilemmas of one's being, but rather to the awareness of oneself as a physical object in the environment. Lower animals, and especially animals that interact with their environments in a relatively localized way (e.g., tactile, gustatory, and olfactory interactions) can operate from a primarily subjective perspective; that is, the world is understood from a perceiver-centered perspective (the world is experienced as centered around the animal, and the animal's actions are experienced as reorienting and reorganizing the surrounding environment). More complex animals, especially those that engage in high-speed, complicated spatial maneuvers (e.g., arboreal monkeys: Povinelli & Cant, 1995), need to have representations of their bodies' positions, orientations, and configurations in space. That is, they require a more objective perspective on the world, in which they understand their own bodies as objects in an independently existing world. Their actions do not so much affect a surrounding subjective universe as affect their body in an objective environment shared by other independent and independently acting objects. Similar considerations apply to animals that coordinate high-speed, spatially distributed group activities in a shared environment (e.g., hunting packs).

Of course, even for these animals, although the planned and experienced ultimate effects of action are understood in reference to an objective environ-

1
Introductions to evolutionary psychology can be found in many recent textbooks, such as Buss (2004) and Gaulin & McBurney (2004).

ment, the subject-centered perspective is not irrelevant (since the immediate effect of most actions is to cause some bodily change). Therefore, higher animals need to coordinate several reference frames, including at least world-centered, local-environment-centered, body-centered, and head-centered frames. This is a complicated constraint satisfaction problem, which under normal conditions is seamlessly and unconsciously solved by neural information processing. Autonomous robots that are intended to operate under similar conditions (high-speed motion, spatially distributed coordination) similarly require this kind of self-awareness in order to control their motion through a shared, objective environment. Therefore they too need to represent their positions, orientations, and configurations with respect to multiple reference frames, and to be able rapidly maintain the mutual consistency of these representations.

Another function of consciousness, in humans at least, is *metacognition*, that is, awareness and knowledge concerning the functioning of one's own nervous system. For example, you may be aware that you are less coordinated when you are tired, that you have a bad memory for faces, or that you act rashly when angry. This is, of course, another form of self-objectification, and may be just as valuable in some autonomous robots as it is in humans.

An additional level of self-objectification facilitates reasoning about the consequences of one's actions. The effect is to step back, view oneself as though another person, and come to an understanding about how one's own psychological processes lead to outcomes that are either desirable or undesirable (either from one's own or a wider perspective), using the same cognitive processes that are used for understanding other people's psychological states and behavior (e.g., neuronal "mirror cells": Rizzolatti & Craighero, 2004). For example, you may recognize that undesirable consequences follow from hitting people when you are angry with them. In this way we acquire a level of executive control over our psychological processes (an important function of *ego-consciousness*, according to psychologists, e.g., Stevens, 2003). For example we can learn (external or internal) stimuli that should trigger more deliberate ("voluntary") control of behavior.

Similar considerations apply to autonomous robots that implement higher-level learning and control of behavior. Such a robot may need to control the operation of its lower-level behavioral programs on the basis of reasoning about the consequences of its own actions (viewed objectively) in its environment.[2] Such control may be implemented through discursive reasoning as well as through analog simulation (e.g., via mirror cells).

I should remark that the account of consciousness presented here is consistent with that of many psychologists (e.g., Stevens, 2003), who observe that consciousness is not the central faculty of the psyche around which all the others orbit (a point also stressed by Jung, 1969b, § 391). Rather, consciousness is a specialized module that is dedicated to handling situations that go beyond the capabilities of other cognitive modules (sensorimotor modules, automated behavioral programs, etc.). We expect conscious robots, like animals, to perform many of their operations with minimal engagement of their conscious faculties. Consciousness is expensive and must be deployed selectively where it is needed.

In summary, we have seen from this review of the functions of consciousness in humans and other animals that many of these functions may also be useful in autonomous robots. Fortunately, applying these ideas in robotics does not raise any great, unsolved philosophical problems. That does not mean that

they are solved, or easy to solve; only that the "less hard" – but still difficult! – methods of neuroscience and neuroethology can be applied to them. As we gradually come to understand the neuronal mechanisms implementing this *functional consciousness* (or *access consciousness*: Block, 1995), we may begin to apply them in robotic design so that our robots can benefit from them as well (and thus exhibit functional consciousness as well).

Intentionality

Intentionality is an issue closely related to consciousness, but not identical to it, so it will be worthwhile to discuss briefly intentionality in artificial agents, such as robots.

Intentionality may be defined as the property by which something (such as a linguistic expression) is *about* something else.[3] Therefore, it is through its intentionality that something is *meaningful* and has *content*. When applied to consciousness, intentionality is the property through which consciousness has content, for consciousness is always consciousness *of* something, as Brentano (1925, p. 89) stressed.[4] Of course, most of the data in a computer's memory is about something – for example, an employee's personnel record is about that employee – but we would not say that the data is meaningful to the computer or that the computer understands it. The intentionality of the data in the computer is dependent upon its meaningfulness to us. Therefore, philosophers have distinguished the *derived intentionality* (of ordinary computer data, books, etc.) from the *intrinsic* (or *original*) *intentionality* of our conscious states, memories, communication acts, etc. (Dennett, 1987, pp. 288–9, Haugeland, 1997, pp. 7–8).

Robots store and process many kinds of data. Much of it will have only derived intentionality, because the robots are collecting and processing the data to serve the needs of the designers or users of the robots. However, in the context of robot consciousness, we are more concerned with intrinsic intentionality, with the conditions under which a robot's internal states and representations are meaningful to the robot itself (and, hence, we could say that the robot *understands*). Each of us can determine by introspection if *we* are understanding something (which is the basis of the Chinese Room Argument), but this will not help us to determine if a robot is understanding, so we must use a different strategy to answer questions about intrinsic intentionality in robots.

The investigation of intrinsic intentionality in non-human agents is a complicated problem, which cannot be addressed in detail here.[5] Fortunately, ethologists have had to deal with this problem in the context of animal communication and related phenomena, and so we may learn from them. For example, animals may act in many ways that influence the behavior of other animals, but which of these actions should be considered communication? One animal, for instance, may sharpen its claws on a tree, and another animal, when it sees

2
This can be viewed as a specialized, high-level application of Brooks' (1987) *subsumption principle*.

3
See, for example, Blackburn (1994, p. 196), Gregory (1987, p. 383), Gutenplan (1994, p. 379), and Searle (1983, p. 1).

4
The philosophical concept of intentionality, in the sense of "aboutness" or meaningfulness, should be carefully distinguished from the ordinary idea of "intention" as purpose or goal.

5
For a fuller discussion see MacLennan (1992, 2006) and MacLennan & Burghardt (1993).

the marks, may go in a different direction. Was this communication, or was it a non-communicative event in which the behavior of one animal indirectly influenced that of another? We would like to be able to determine whether the *purpose* of the first animal's action was to influence the behavior of other animals (e.g., Grice, 1957), or whether that was merely an accidental consequence of its action (but not its purpose).

As we have seen, the best way to understand purpose in a biological context is to look to a behavioral adaptation's selective advantage, or lack thereof, in a species' environment of evolutionary adaptedness (EEA). In this way, communication can be defined as an action that, in the EEA, has the statistical likelihood of influencing the behavior of other animals in such a way as to increase the inclusive fitness of the communicator (that is, the selective advantage of the communicator or its group) (Burghardt, 1970). In a similar way we can approach the intrinsic intentionality of other meaning-bearing states or representations in any agent (animal, robot, etc.). To a first approximation their meaning is grounded in their relevance to the survival or well being of an individual agent, but it is more accurate to ground meaning in the agent's inclusive fitness, which takes account of its selective advantage to the agent's group. Of course, the meanings of particular states and representations may be only loosely and distantly correlated to inclusive fitness, which nevertheless provides the ultimate foundation of meaning.

Perceptual-behavioral structures and their associated representations that have a significant genetic component need to be interpreted in reference to the EEA. Behaviors and representations that have no selective advantage in an animal's current environment (e.g. hunting behavior in a captive or domesticated animal) may have a meaning that can be understood in the context of the EEA. This does not imply that an agent's internal states and behavior have no meaning in other environments, but only that the meaning of innate perceptual, behavioral, and cognitive structures should be interpreted in the context of the EEA (for it is *that* environment that defines their purposes and has given them their primary meaning).

Can artificial agents, such as robots, exhibit intrinsic intentionality? *Synthetic ethology* offers a methodology by which such questions can be addressed (MacLennan, 1990, 1992, 2006; MacLennan & Burghardt, 1993). The goal of synthetic ethology is to permit the scientific investigation of problems relating to the physical processes underlying mental phenomena by studying synthetic agents in "synthetic worlds", which are complete but very simple, and so permit the conduct of carefully controlled experiments. For example, in one series of experiments beginning in 1989 we used synthetic-ethology techniques to demonstrate the evolution of communication in a population of simple machines (MacLennan, 1990, 1992). We showed that if the machines are able to modify and sense a shared environment, and if there is selective pressure on cooperative behavior (which could be facilitated by communication, but could also occur without it), then the machines will evolve the ability to communicate. The signals exchanged by these machines are meaningful *to them* because, in their EEA, these signals are relevant to the machines' continuing "survival" (as organized structures). As observers we can monitor their behavior and infer the meaning of their communication, but in this case *our* understanding is derived, whereas *theirs* is intrinsic.[6]

Such experiments help us to articulate the differences between consciousness and intentionality, for although these simple machines can exhibit intrinsic intentionality in their communication, they are not conscious (or even alive). In itself, this should not be too surprising, for very simple animals, such as

bacteria, communicate with each other and have internal states that represent their environment (e.g., Dretske, 1985, p. 29; Dunny & Winans, 1999); their internal states and signals have intrinsic intentionality, although they do not exhibit consciousness in the sense that I have used it hitherto.

With this background, we can address the question of intrinsic intentionality in robots and its relation to consciousness. Certainly, truly autonomous robots need to be concerned with their own survival: for example, they need to be able to find energy sources (e.g., sunlight, fuel), to repair themselves (to the extent possible), to extricate themselves from dangerous situations (e.g., stuck in mud or sand), to avoid natural threats (e.g., weather, unsafe terrain, curious or predatory animals), and perhaps (for military robots) to evade, escape, or neutralize hostile agents. Functions such as these, relevant to the robot's continued existence qua robot, provide a foundation of intrinsic intentionality, which grounds the robot's cognitive states, for they are meaningful *to the robot*.

Such functions contribute to an *individual* robot's fitness, but there are other circumstances in which it would be advantageous to have a robot sacrifice its own advantage for the sake of other robots. For many purposes we need cooperative groups of robots, for which the collective fitness of the group is more important than the success of its members. Indeed, these same considerations apply to robots that define their group to include (certain or all) human beings or other groups of animals, for whom they may sacrifice their own advantage. In all of these "altruistic" situations, group fitness provides an expanded foundation of intrinsic intentionality.

Finally, for some applications it will be useful to have self-reproducing robots; examples include applications in which robots might be destroyed and need to have their numbers replenished, and situations in which we want to have the number of robots adapt to changing conditions (e.g., expanding or contracting with the magnitude of the task). If the robots reproduce sufficiently rapidly (which might be the case, for example, with genetically engineered microorganisms), then we must expect microevolution to take place (for the inheritance mechanism is unlikely to be perfect). In these situations, intrinsic intentionality will emerge from the inclusive fitness of the members of the evolving population in the environment to which it is adapting, just as it does for natural populations. Therefore we can see that under a wide variety of circumstances, the conscious states of robots will have intrinsic intentionality and thus genuine content; their consciousness will be consciousness *of* something, as it must be. (I mention in passing that emotions, which have many important connections to consciousness, are important in all these applications of autonomous robotics.)

3. The Hard Problem

Why It Is Hard?

Having discussed the "less hard" problems of animal and robot consciousness (which are certainly hard enough to keep us busy for many years!), I will turn

6
Recent work on computer-based investigations of the evolution of language and communication can be found in Cangelosi & Parisi (2001) and Wagner, Reggia, Uriagereka & Wilkinson (2003); unfortunately the latter incorrectly classify our own experiments as nonsituated communication. See MacLennan (2006) for a more detailed discussion of these issues.

to the Hard Problem in the context of human consciousness and as it may arise in the context of robot consciousness.

The Hard Problem, which addresses the relation of our ordinary experience of subjective awareness to the scientific world-view, is arguably the principle problem of consciousness (MacLennan, 1995), and so it will be worthwhile to say a few words about what makes it so hard.[7] The root of the problem is the unique epistemological status of *phenomenal consciousness* (Block, 1995), for conscious experience is the *private* and *personal* ground of *all observation*, whereas traditionally science has been based on *specific observations* that are *public* and, in this sense, *non-personal*. We are dealing with several interrelated epistemological issues.

First, in order that its conclusions may be generally applicable, science strives to be a *public* enterprise, and so it is based on publicly validated observations, whereas the experience of conscious awareness is inherently *private*. (Verbal accounts of conscious awareness can, of course, be public, but assuming that they are veridical begs the question of the Hard Problem.) On the other hand, it is important to recognize that all observation is ultimately private, and that in consciousness studies, as in the more developed sciences, a body of public facts can emerge as a consensus from the reports of the private experiences of trained observers of differing theoretical commitments. I will address the sort of training required below.

Since the goal of science is public knowledge (knowledge true for all people), science seeks to separate the observer from the observed, for it wants its conclusions to be founded on observations that are independent of the observer. This is not feasible when the object of scientific investigation is conscious experience, for consciousness constitutes the state of *observation*, comprising both the observer and the observed, the fundamental relation of *intentionality*, as described by Brentano (1995, p. 89) and Husserl (1931, p. 34): *intentio* (stretching, direction, attention) towards an object. Consciousness is the vector of intentionality extending from the observer to the observed, and so it involves them both essentially.

Further, science ordinarily strives to separate the individual, *subjective* aspects of an observation (e.g., felt warmth) from the *objective* aspects (e.g., measured temperature), about which it is easier to achieve a consensus among trained observers. However, in the Hard Problem the individual, subjective aspects are of central concern. Also, science normally takes a *third-person* perspective on the phenomena it studies (*it, he, she* is, does, etc.), whereas the experience of conscious awareness is always from a *first-person* perspective (*I* feel, perceive, remember, etc.). Indeed, the Hard Problem addresses the question of why, in a fundamental sense, there even *is* a first-person perspective.

Indeed, it might seem that an *objective* science of *subjective* experience is impossible, a contradiction in terms, but this impression results from a confusion of terminology. Here I use "subjective" and "objective" to refer, respectively, to private, first-person experience and to public, third-person observation. Often, however, we understand "objective" to mean "unbiased or factual" (and therefore *good*), and "subjective" to mean "biased or distorted" (and therefore *bad*). As Searle (1992, p. 19) suggests, progress on the mind-body problem has been impeded by a pun! Of course, the descriptive and evaluative usages of these terms are not unrelated, but our goal here is objective (i.e., unbiased, factual) knowledge of subjective (i.e., first-person, private) phenomena.

The inherently first-person, subjective character of conscious experience makes it resistant to the ordinary reductive patterns of science, for it is the third-person, publicly observable aspects of phenomena that are most amenable to reduction to more fundamental physical processes. For example, once the private experience of felt warmth has been separated from the public measurement of temperature and heat, the latter can be reduced to more fundamental physical properties (mean kinetic energy of molecules). Indeed, although third-person objects, properties, and processes can be reduced to other third-person objects, properties, and processes, it is a category mistake to attempt to reduce first-person phenomena to the third-person objects, properties, or processes. Nevertheless, there is a kind of reduction that is applicable to subjective phenomena, as explained below.

Observing Consciousness

The unique epistemological status of conscious experience makes it difficult to investigate by scientific means, but not impossible; here I will summarize the approach that I have advocated (MacLennan, 1995, 1996a, in press). First I will address the question of how we can observe consciousness (i.e., look *at* it), when all observation is by means of consciousness (i.e., looks *through* it). An analogy will make the approach clear.

The image formed by a camera must pass through the camera's aperture; in this sense, we can take a picture of some object (analogous to the content of consciousness), but we cannot take a picture of the aperture itself (analogous to observing consciousness). Nevertheless, it is possible to investigate the aperture, because it affects the image in systematic ways (e.g., brightness, diffraction, focus, depth of field). In particular, we can investigate changes that occur with systematic variation of the aperture. In this way, characteristics of the aperture may be separated from the specifics of the image. In some cases these observations are facilitated by the use of a simple object, such as a point light source or a *ganzfeld* (homogeneous field), which reveals some characteristics of the aperture, but obscures others that are peculiar to complex images.[8] Therefore, armed with insights gained from simple images, it is also necessary to explore the affects of the aperture on complex images. In any case, many of characteristics of the aperture will be unapparent to the naive observer, but with training they are uncovered, and provide the basis for a body of public facts. Trained investigators will be able to explore the affects of varying the aperture on all images, and thereby discover objective relationships.

This analogy suggests an approach to observing consciousness. Although consciousness cannot be separated from its content, trained observers can separate aspects of the conscious state that depend more on its content from those that depend on consciousness itself. As in the camera analogy, investigation may be facilitated by conscious content that is simple in structure,

7
A fuller discussion can be found in Chalmers (1995, 1996), MacLennan (1995, 1996a), and Searle (1992, chs. 4–5).

8
The behavior of a linear system (such an optical system) is completely characterized by its *impulse response*, which describes its behavior when its input is an idealized point source or impulse. Linear systems can also be characterized by a kind of *ganzfeld*: a "white noise" input signal in which all wavelengths are equally represented. (See any textbook on linear systems analysis.)

as occurs in contemplation and meditation (e.g., emptiness, one-point concentration). More generally, consciousness can be investigated in laboratory situations that attempt to control its content. However, as our analogy suggests, such approaches reveal only some characteristics of consciousness while obscuring others. Therefore, it is essential also to investigate ordinary, everyday conscious states, as well as altered states that accentuate particular characteristics.

It will be apparent that specialized training and experience are necessary to observe the relevant phenomena, as they are in all sciences, but especially in the scientific study of consciousness. Experimental phenomenology and phenomenological psychology (e.g., McCall, 1983), since they directly address the structure of phenomena (conscious experience), seem to provide the best foundation. Ihde (1986) shows how systematic variation of simple phenomena can help to reveal the structure of consciousness.

Neurophenomenology

In its literal sense, a *phenomenon* (Greek, *phainomenon*) is anything that appears (*phainetai*) in consciousness; among the kinds of phenomena are perceptions, thoughts, recollections, plans, intentions, volitions, desires, fears, anticipations, and hallucinations. But phenomena are not independent; they exist in interrelationships of sequence and possibility. This network of actual and potential phenomena constitutes a *phenomenal world*. *Phenomenology* is fundamentally the study of the structure of phenomena, that is, of the invariant structure of phenomenal worlds (the structure independent of individual variation). Since an adequate scientific theory of consciousness must account for qualia and their integration into a phenomenal world, phenomenology is fundamental to the consciousness research.

By using phenomenological techniques, investigators can avoid an overly superficial perspective on phenomena, often based on an *a priori* theoretical commitment. Consider a well-known example from Husserl's *Cartesian Meditations* (1960, §§17–19). Suppose someone rotates an ordinary die within my view. What would be an accurate phenomenological description of my experience? One might suppose that it might be an account of neutral visual data in terms of changing configurations of black ellipses in white parallelograms, but this is not an accurate description. In fact, I do not experience abstract ellipses and parallelograms; I experience a rotating cube marked black spots. Indeed, since I am acquainted with dice, I will experience a rotating die. The recognition of this familiar three-dimensional object is an aspect of the phenomenon. Furthermore, the phenomenon is not confined to the instantaneous present; by means of short-term memory it extends into the recent past (*retention*, in Husserl's terms), and by means of anticipation it extends into the near-term future (*protention*); this actual and potential sequential structure gives the phenomenon its temporal unity (e.g., Husserl, 1973, §23). There are also non-visual anticipations and associations, for we expect the die to have a certain hardness and weight. Violations of certain expectations (e.g., discovering that the die has no back, or is extremely heavy) lead to a kind of breakdown, and a change in our intentional relation to the object. The structure of the die phenomenon is not limited to perception, but has psychological and social aspects. For example, I will associate the die with games and gambling (and whatever connotation that may have for me), and I may even experience

the displayed die as an invitation to some sport. All this is part of the rich phenomenology of so simple a thing as a die.

Accurate phenomenology depends on awareness and investigation of all aspects of the phenomena, a skill that requires significant phenomenological training. The technique is far from naive introspectionism, which is, indeed, naive.

If we want to solve the Hard Problem, that is, to understand the relation of consciousness to physical processes, we cannot rely on phenomenology alone, but must integrate phenomenological observation with neuroscientific theory and experiment. Each domain of investigation may contribute to the other. For example, we know that rapid and slow motions are processed differently in the brain (Weiskrantz, 1995), which should alert us to look for corresponding phenomenological differences. On the other hand, the phenomenological subtleties of color (discussed later) imply corresponding neurological processes. Therefore, the scientific investigation of consciousness must be, in a broad sense, *neurophenomenological*.[9]

Protophenomena

Neurophenomenological Reduction

The value of reduction is that it allows us to understand complicated systems better by relating them to simpler systems. (Reduction is most fruitful when it does not limit itself to understanding how the parts constitute the whole, but also considers the role of the whole in the constitution of the parts; this is especially the case in the biological, psychological, and social sciences.) Therefore, although a reduction of the subjective to the objective is fundamentally impossible, we can accomplish a reduction of the subjective to the subjective (that is, a reduction of subjective phenomena to their subjective constituents) and, further, correlate this subjective reduction to a parallel reduction, in the objective domain, of neuropsychological processes to their constituent biological and physical processes.

Reduction in the subjective domain can be accomplished by observers trained in phenomenological procedures, which allow them to arrive at a consensus concerning the structure of conscious awareness as experienced by all people. (There is already a considerable body of results, in the psychological literature as well as the phenomenological literature.) As we've seen, insights and results from each of these domains – which we may call the phenomenological and the neurological – can suggest hypotheses and otherwise guide the investigations of the other.

As a first step we can attempt a *qualitative reduction*, essentially a "separation of variables", of phenomena on the basis of sensory modality; for example visual phenomena of all sorts (perceptions, memories, etc.) can be separated from auditory phenomena. Thus the conscious state is decomposed into phenomena of different kinds. Even here, however, we must beware of oversimplification, for neurological research has shown that some neurons is auditory

9
The term was coined, apparently, by Laughlin, McManus & d'Aquili (1990).

10
On auditory cortex, see Pribram (1991, p. 81, citing Bridgeman, 1982), Pribram, Spinelli & Kamback (1967), and Bavelier & Neville (2002); on visual cortex, see Calvert et al. (1997).

cortex respond to visual stimuli, and conversely neurons in visual cortex can respond to auditory stimuli, thus facilitating face-to-face communication.[10] This suggests that ostensibly visual phenomena are not purely visual, nor are supposed auditory phenomena purely auditory, and it is reasonable to suppose that the same mixture occurs among other sensory modalities (as suggested also by the die example). Therefore, a qualitative reduction can be at best approximate, as we should expect from both pure phenomenology and evolutionary psychology (i.e., visual dominance).

In contrast to a qualitative reduction, which decomposes phenomena on the basis of kind, it is possible to perform a *quantitative reduction*, which decomposes phenomena on the basis of size. This approach is suggested by philosophical considerations, but also by neuroscience. In particular, *topographic maps* and other *computational maps* are ubiquitous in the brain (Anderson, 1995, ch. 10; Knudsen, du Lac & Esterly, 1987). For example, in sensory areas the dimensions of a stimulus are systematically mapped onto cortical regions. The most familiar example is the *somatotopic map* in somatosensory cortex, in which cortical location corresponds to bodily location, but there are similar body maps in motor areas. In visual areas there are *retinotopic maps*, in which neural location corresponds systematically to retinal location. The mapped dimensions of the stimulus can be more abstract. For example, in auditory cortex there are *tonotopic maps*, in which neural location corresponds to frequency, and in bat auditory cortex, echolocation is aided by maps encoding Doppler shift (Suga, 1985, 1989).

Although there is much that we do not know about neural representation, these examples suggest that many representations can be decomposed into elementary units (i.e., individual neurons, or small groups of them),[11] that are essentially similar in function and distinguished only by their location in some computational map. Furthermore, at least in primary sensory areas, it has been possible to relate activity in these neurons to elementary constituents of stimuli (e.g., pressure on a particular patch of skin, light of certain wavelengths on a particular retinal location), the *receptive fields* of the neurons. This is all in the neurological domain, but we can perform a parallel reduction in the phenomenological domain, for we are aware that, for example, visual phenomena have parts, such as our experiences of color at different locations in the visual field (an observation that applies to visual hallucinations as much as to ordinary perception). The elementary components of a phenomenon, then, would correspond to the smallest units of the corresponding neural representation (presumably, activity in individual neurons, but other possibilities are considered below).

Thus, neurologically-informed phenomenological reduction (which we may call *neurophenomenological reduction*) suggests that it may be fruitful to understand conscious experience in terms of *protophenomena*, which are theoretical entities hypothesized as the elementary constituents of phenomena. We further hypothesize that each protophenomenon has an *intensity* (a sort of fundamental *quale*) representing its presence in consciousness (e.g., experienced pressure on a patch of skin, experienced brightness of a patch of color in the visual field). This intensity is the subjective experience corresponding to neural activity in the neural structures associated with a protophenomenon (its *activity site*). (I will discuss activity sites in more detail below.)

The simplest kinds of protophenomena are similar to elementary sense data (such as "red-here-now"). For example, if we consider visual experience, we can think of it as constituted of tiny patches of color and brightness, much like

pixels, at various locations in the visual field.[12] However, neuroscience suggests that ostensibly visual protophenomena will also have an auditory aspect, and vice versa. Furthermore, protophenomena are not limited to elementary sense data, but also include the elementary constituents of more complex phenomena, including expectations, moods, feelings, recollections, imaginations, intentions, and internal dialogues. In any case, neurophenomenology suggests that protophenomena are very small compared with phenomena, and one's conscious state might comprise 10 to 100 billion protophenomena (the number of neural activity sites associated with protophenomena). Protophenomenal interdependencies are also much more complex than suggested by the notion of elementary sense data (as is discussed below), so we must beware of an oversimplified or superficial understanding of protophenomena. Indeed, as neurons both sense their cellular environments (via chemical receptors) and act on their environment (by generating action potentials), so most protophenomena have an active character, in that their presence in consciousness conditions the presence or absence of other protophenomena.

We identify one's phenomenal world with the totality of their protophenomena, but this may seem to lead to a "jaggedness" or "grain" problem (Chalmers, 1996, pp. 306–8) in the absence of some additional factor to unify the protophenomena into a whole, but this is not the case. Consider a macroscopic object such as a chair; it is a whole because its constituent atoms are bound together, so that their macroscopic motions are coherent. Similarly, as will be explained, the intensities of protophenomena are mutually interdependent, and a phenomenon is no more than the coherent activity of masses of protophenomena. So also, the unity of consciousness is a consequence of the unity of the nervous systems (see "The Unity of Consciousness" and "The Unconscious Mind" below).

Ontological Status

Since, in a philosophical context, a phenomenon is anything that appears in consciousness, phenomena are, by definition, observable (indeed, from a first-person perspective). Paradoxically, protophenomena, which are the elementary constituents of phenomena, are not, in general, observable. This is because under normal circumstances protophenomena are experienced only as parts of whole phenomena, which typically comprise millions of protophenomena (as will be explained below), so that a change in one protophenomenon would rarely be noticed (i.e., cause one to behave differently). As an analogy: the change of one pixel in a high-resolution image is unlikely to have any practical effect. Similarly, changing one molecule of a macroscopic object (such as a chair) is unlikely to have a noticeable effect. Conversely, just as bound and coherently moving atoms constitute a macroscopic object, so bound and coherently varying protophenomena constitute a phenomenon present in consciousness (protophenomenal interdependencies are discussed later). We may say that the protophenomena constituting a phenomenon have

11
Such as microcolumns, containing perhaps eleven neurons (Jones, 2000).

12
The primary protophenomena of visual experience appear, in fact, to be more complex then pixels; psychophysical evidence suggests their brightness profiles are more like spatiotemporal Gabor wavelets (Pribram, 1991); see also MacLennan (1991) for a survey.

essential subjectivity, but are not themselves phenomena. That is, *proto*phenomena are not the same as little phenomena.

The apparent unobservability of protophenomena raises questions about their existence. In our current state of knowledge it is perhaps best to view them as *theoretical entities*, which means they are postulated for their explanatory value in the theory and are validated by their fruitfulness for scientific inquiry (Hempel, 1965, pp. 177–9; Maxwell, 1980, pp. 175–84). Their ontological status is comparable to that of atoms during the nineteenth and early twentieth centuries, when they could not be observed directly. Physicists might have differed (especially in the nineteenth century) about whether atoms *really* exist, but they all agreed on the scientific value of atomic theory. (In contemporary physics, quarks and strings are unobserved theoretical entities.)

There are other possibilities. For example, protophenomena might be emergent properties of sufficiently large or complex brains, but this possibility does not necessarily imply that they are not real or that there is some critical neural mass below which they do not exist. Again, an analogy will help. Sound is a compression wave in a medium such as air, and such a wave can be understood by assigning a pressure to each point in a volume of space. We know this is a mathematical fiction, since air is composed of discrete molecules, and it makes little sense to talk of the pressure of one or two molecules or even of a small number of them. Nevertheless, sound and pressure distributions are perfectly objective properties of macroscopic volumes of air. So also we may find it is meaningful to talk of protophenomena only in the context of macroscopic neural mass.

Activity Sites and Protophenomenal Intensity

Parallel reduction in the phenomenological and neurological domains leads to the conclusion that there are *activity sites* in the brain corresponding to the protophenomena, and that some kind of physical process at an activity site corresponds to the intensity (strength) of the corresponding protophenomenon in conscious experience. It is important to understand that a protophenomenon and its activity site are two mutually irreducible aspects of a single underlying reality (and thus protophenomena theory is a kind of *double-aspect monism*).[13]

Unfortunately, I do not believe that we can identify the activity sites at this time. Some reasonable possibilities include synapses and neural somata, in which cases the intensity of the associated protophenomenon might correspond to neurotransmitter flux, bound neurotransmitter receptors, or membrane potential. Following Sherrington, who said, "Reflex action and mind seem almost mutually exclusive–the more reflex the reflex, the less does mind accompany it", Pribram has argued that consciousness is associated with the graded electrical activity in the dendritic trees of neurons, rather than with all-or-nothing action potential generation.[14] On this basis we would expect synapses to be the activity sites and protophenomenal intensity to be correlated with neurotransmitter flux, bound receptors, or pre- or postsynaptic membrane potential.[15] A related possibility is that neural somata are the activity sites, and that intensity corresponds to somatic membrane potential, which is also graded; other possibilities are considered below in "Consequences and Issues". In any case, these are all scientific questions, which can be addressed empirically.

As previously discussed, a protophenomenon has a degree of presence in consciousness, which we call its *intensity* (think of the brightness of the red-here-now for a concrete example), and we hypothesize that this intensity is correlated with some physical property of the activity site, for example membrane potential, neurotransmitter or ion flux, or the number of bound receptors. The simplest hypothesis is that protophenomenal intensity is simple, nonnegative, scalar quantity (representing degree of presence), but there are other possibilities. For example, protophenomena associated with different neurotransmitters might have different kinds of intensities, and consequently a different experienced presence in consciousness; this is an empirical question that can be answered by experimental phenomenology.

Protophenomenal Dependencies

An important issue is what distinguishes, for example, a protophenomenon for "red-here-now" from one for "middle-C-here-now", that is, what gives protophenomena their qualitative character? The parallel question in the neuroscience domain suggests an answer, for neurons in visual cortex, for example, are not essentially different from those in auditory cortex. Certainly the sensory receptors are different, but even in the sense organs there is no important difference between, for example, a cone responding to certain optical wavelengths at one place on the retina from those with the same response at other places. Rather, the *structure* of the sensory world is defined by the interconnections among neurons. For example, the spatial structure of vision is defined by patterns of connections that cause neurons to respond to edges, lines, center-surround patterns, and other spatial structures.

Protophenomena seem to be organized according to similar principles. That is, the time-varying intensities of protophenomena are correlated with each other in accord with quantifiable *protophenomenal dependencies*; in principle these correlations can be described by differential equations (MacLennan, 1996b, in press). That is, the intensity of each protophenomenon is a complicated function of the recent intensities of thousands (or tens or hundreds of thousands) of other protophenomena, as well as of *extrinsic variables*, that is, of variables external to the phenomenological domain. As a consequence, the phenomenal world is not causally closed, but the protophenomenal dependencies constrain the possibilities of change in conscious state, subject to the extrinsic variables and other influences discussed below.

It is reasonable to say that protophenomena have no *qualities* of their own; they have only their intensities (which are *quantities*); protophenomena have qualities only by virtue of their interdependence with other protophenomena.

13
More specifically, protophenomena theory is an example of what Chalmers (2002) calls *type-F monism*, which is in the heritage of Russell (1927). Jung's phenomenological psychology led him to similar conclusions: "psyche and matter are two different aspects of one and the same thing" for "the biological instinctual psyche, gradually passes over into the physiology of the organism and thus merges with its chemical and physical conditions" (Jung, 1960, § 418, 420). See also Jung & Pauli (1955) and Stevens (2003, pp. 79–88).

14
See Miller, Galanter & Pribram (1960, pp. 23–4) and Pribram (1971, pp. 104–5, 1991, pp. 7–8).

15
This possibility is developed mathematically in MacLennan (1996b, Appendix); see also MacLennan (1999b).

Therefore, qualia are emergent properties in a phenomenal world structured by protophenomenal dependencies; that is, this is essentially a *structuralist* theory of qualia.

Phenomena are experienced "out there" – in our physical bodies or in the space around them – and only rarely inside our heads, where cortical neural activity occurs. We see objects (and even hallucinations) around us, not in our visual cortices, and we feel pains in our fingers or toes, not in our somatosensory cortices. Why do we experience activity in one cortical neuron as a pain in a finger, and in another as a pain in a toe? Topographic maps in the brain suggest an answer, for spatial relations in the map mirror spatial relations among the stimuli.[16] But spatial proximity in the cortex is not in itself the primary factor (although diffuse electrical and chemical effects are possible, and the brain's EM field may play a role); rather, the key factor is that in topographic maps nearby neurons are more likely to be connected than are more distant neurons. Interactions among nearby neurons generate a topology (an abstract system of neighborhood relationships), which creates the phenomenal space into which our experiences are projected. Since protophenomenal dependencies correspond to physical dependencies among their activity sites, protophenomenal dependencies define the topology of the phenomenal world, which is a major aspect of its phenomenology, that is, of the possible structure of phenomena (MacLennan, 1999b).

Recent experiments by Sur (2004) support the dependence of phenomenal quality on neural interconnection. Retinal axons in newborn ferrets were induced to project into auditory cortex (area A1, via the thalamus) in one hemisphere, but projected to their normal targets in visual cortex (V1) in the other. As a consequence, the auditory cortex that received retinal input self-organized into orientation maps like those in primary visual cortex. Furthermore, and most significantly, neurophysiological and behavioral tests implied that the ferrets were *experiencing visual perceptions* in their "rewired" auditory cortices.

Phenomenological Change and Closure

Protophenomenal dependencies determine the structure of one's phenomenal world and therefore one's possible conscious states, but the structure of this world is not fixed. First, short- and long-term learning alters the connections between activity sites and therefore the effects that each has on the others. Correspondingly, learning affects the interdependencies among the intensities of protophenomena, altering the possibilities and probabilities in the sequence of possible phenomenal states. As a result, protophenomenal intensities may become more tightly coupled, so that they vary coherently, constituting a phenomenon proper. Thus, what was previously unmanifest can become apparent in consciousness.

In this article I will not address phenomenological changes that take place during individual development (e.g., as a result of several cycles of neural proliferation and programmed cell death), but will focus on plasticity in the adult. It is now well established that in adults, neurons can make new connections (Shepherd, 1994, pp. 222–3), and there is accumulating evidence for new neuron growth in the hippocampus and perhaps in other areas.[17] If any of these processes generate new activity sites, then there will be new protophenomena to accompany them, effectively expanding the degrees of freedom of the phenomenal world.

As mentioned above, phenomenal worlds are not causally closed; protophenomenal dependencies do not completely determine the dynamics of a phenomenal world. The principal non-phenomenological causes are the extrinsic variables corresponding to sensory inputs. However, other physical processes can also affect the phenomenal world. For example, sickness and alcohol or other mind-altering substances can temporarily affect protophenomenal dependencies. More permanent changes to one's phenomenology can result from strokes, brain tumors or injuries, Alzheimer's disease, and the like.

The incompleteness of phenomenological causality might seem to imply that the phenomenal world is ultimately epiphenomenal, and that protophenomenal theory is unnecessary in the presence of a (presumably) causally complete physical theory. However, this familiar perspective ignores the Hard Problem, since it does not address phenomenal consciousness at all; that is, a substantial body of evidence remains unexplained. In contrast, the protophenomenal approach allows a reduction within the subjective domain, the correlation of elementary subjectivity with physical processes, and the eventual integration of consciousness into the scientific worldview.

Consequences and Issues

Inverted Qualia

The idea of a color spectrum inversion dates back at least to Locke's 1690 *Essay Concerning Human Understanding* (e.g., Hardin, 1988; MacLennan, 1999a; Nida-Rümelin, 1996; Palmer, 1999). Is it possible that I experience phenomenal redness when I perceive short wavelengths (normally experienced as violet), and vice versa? Neurophenomenological reduction and the protophenomenal approach provide means for answering these questions empirically.

To illustrate the approach I will begin with a simpler problem: an auditory spectrum inversion. It might seem conceivable that I experience as a phenomenal high pitch the same sound frequencies that you experience as a low pitch, and vice versa, but this apparent possibility rests on a superficial phenomenology of pitch, which can be exposed by systematic variation of the phenomena. On one hand, if we gradually increase the subjective pitch of a sound, we will discover a limit beyond which we cannot go (in perception or perceptual imagination). On the other hand, if we gradually decrease subjective pitch, we find that it comes to be experienced more a rhythm and ultimately as a periodic variation of loudness. (To be more specific, frequencies above, say, 100 Hz are experienced as pitch, whereas those below about 10 Hz are experienced as rhythm; intermediate frequencies are experienced in a mixed way.) Thus experimental phenomenology demonstrates that our experience of low pitches is distinguished from that of high pitches in that the former are inherently continuous with our experiences of rhythm and loudness.

16

Additional evidence comes from "referred pain", a medical condition in which pain in one part of the body is transferred to another part that is not nearby in the body, but whose cortical maps are adjacent. This may occur, for example, because of cortical remapping after loss of a body part (e.g., Karl, Birbaumer, Lutzenberger, Cohen & Flor, 2001).

17

See, for example, Gould, Reeves, Graziano & Gross (1999) and Rakic (2002).

This phenomenological analysis is reinforced by neuroscience, for higher frequencies are mapped spatially in auditory cortex in tonotopic maps, which limit the representable frequencies (in perception but also imagination) at both the high and low ends. However, at lower frequencies (about 5 Hz and below), nerve impulses become synchronized with the sound waves (i.e., the frequencies are represented temporally rather than spatially), a representation like that of a rhythm (Adelman, 1987, p. 91; Suga, 1994, pp. 299–300; see also Bendor & Wang, 2005). Therefore a more systematic neurophenomenological analysis of sound shows that the alleged spectral inversion is impossible; abnormalities in neural structure would manifest in experience, because the phenomenon of low pitch essentially includes aspects of rhythm, which high pitches do not.

We can apply similar techniques to inversions in visual qualia. The simplest case is an inversion between phenomenal dark (which I'll denote Φ-Dark) and phenomenal light (Φ-Light). In fact, it is impossible because, as Francis Bacon (*Essays*, 3) remarked, "All colors will agree in the dark." In particular, the experience of Φ-Dark does not admit of differing color experiences.

The possibility of a color inversion is suggested by the idea of a linear color spectrum, which is a consequence of inaccurate phenomenology contaminated by knowledge of the physics of light (the linear dimension of wavelength) and analogies with sound (pitch and wavelength). Indeed, prior to Newton color was less likely to be understood as a linear spectrum, but his discovery of the color spectrum established the idea that color is a one-dimensional phenomenon (Gage, 1993). Since Hering's (1878) development of the double-opponent theory of color vision, however, it has been apparent that color has a more complex topology, which is also supported by neuroscience (e.g., De Valois & De Valois, 1988, 1993; Kaiser & Boynton, 1996).

Phenomenal hue has a circular topology structured by two axes between opposing colors, which may be termed (approximately) the yellow-blue and red-green axes (hence, "double-opponent"). The axes are defined by four "unique hues" (unique-yellow, unique-blue, unique-red, unique-green), which are experienced as being unmixed with any other colors. (For example, the experience of unique-green does not have any mixture of blue or yellow in it.) The wavelengths of light that are perceived as these unique hues varies a little from person to person, but they are an essential aspect of the phenomenology of normal human color vision. (Actually, there is no single wavelength that produces the experience of unique-red, but the experience can be created by mixing in blue and yellow wavelengths with red light, so that they cancel each other on the yellow-blue axis; more on this below.) Therefore, at a basic level, human color experience is defined by three axes: yellow-blue, red-green, and light-dark (YB, RG, and LD, respectively), which define a *color sphere*. This structure suggests a number of possibilities for anomalous color vision, for we can entertain exchanges of the opposed colors (e.g., an exchange of yellow and blue) or exchanges of entire axes (e.g., an exchange of YB with RG) (cf. Palmer, 1999).

As previously discussed, Φ-Light and Φ-Dark are phenomenologically different in structure, and therefore cannot be exchanged, so I will focus on the more interesting color exchanges. Indeed the phenomenological differences between Φ-Light and Φ-Dark provide a basis for color phenomenology, since it has been recognized since ancient times (e.g., Aristotle, *De Sensu*, 442a) that yellow and blue are the colors most closely related to light and dark (i.e., white and black); indeed, we may call yellow and blue the chromic analogs

of white and black. Pre-Newtonian linear color theories often understood the colors as intermediaries between white and black, with yellow and blue being closest to the extremes (Gage, 1993). Likewise, modern color researchers refer to the "yellow anomaly", which refers to the fact that phenomenal yellow (Φ-Yellow) is the intrinsically brightest hue. Therefore the experiences of unique-yellow and unique-blue are phenomenologically distinct from each other and from the other colors by virtue of their relations to Φ-Light and Φ-Dark. Furthermore, if someone had an experience of Φ-Yellow when perceiving short-wavelength (blue) light, the abnormality would be detectable (for they would report a "blue anomaly"). Therefore an undetectable YB inversion is impossible.

As remarked, the phenomenological characteristics of Φ-Yellow and Φ-Blue preclude an undetectable exchange of the YB and RG axes, but an inversion of the RG axis might seem possible, since on the color wheel Φ-Red and Φ-Green are both intermediate between Φ-Yellow and Φ-Blue (but on opposite sides of the wheel).

A solution to this problem may be found in the color theory of Goethe (1840), who was a very careful phenomenologist. Although his criticism of Newtonian optics is often viewed as "an embarrassing lapse in the life of an otherwise great man", Goethe had a more accurate account of the phenomenology of color, which in fact complements Newton's account, which was better for the development of physical theory. Indeed, it is not surprising that Goethe the painter would understand color differently than Newton the theoretical and experimental physicist; whereas Newton explored pure wavelengths split out of pure white light by a prism, Goethe investigated naturally occurring color in the sky, clouds, plants, and minerals (also, in some cases, by means of a prism).

Goethe observed that both red and green are experienced as means between the extremes of yellow and blue, but means of a different kind. On one hand, green is a simple intermediate between yellow and blue, similar to both, even though unique-green includes no blue or yellow (Goethe, 1840, § 697). On the other hand, red does not have this relationship but, according to Goethe, by a process of phenomenological "augmentation" (*Steigerung*) of yellow and blue (§ 699–703), one can produce a very pure red (*Purpur*), "like fine carmine on white porcelain" (§ 792). (In this connection it's worth recalling that unique-red is a *non-spectral hue*, that is, it is an experienceable color that cannot be produced by monochromatic light and does not occur in the color spectrum.) Thus we have a basis for the phenomenological distinction of the four unique hues, which is supported by the neuropsychology of visual perception.

This phenomenological analysis is supported by the cross-cultural studies of Berlin and Kay (Berlin & Kay, 1969; Kay & McDaniel, 1978; Saunders & van Brakel, 1997). If a culture has two basic color terms, they are nominally equivalent to *white* and *black*, but have denotations closer to *warm-bright* and *cool-dark*, effectively Φ-Yellow and Φ-Blue. If they have a third basic color term, it is approximately equivalent to *red*, and a fourth is *green*. Similarly, Goethe classifies red as the third primary color (after yellow and blue), and makes green the first secondary color.

Therefore, color phenomenology, the neuropsychology of color vision, and studies of cross-cultural color categorization all imply that the four unique hues are phenomenologically distinct, that each has an individual character, and therefore that anomalies in color vision would be detectable. We can con-

clude that undetectable color inversions are impossible. Furthermore, these neurophenomenological investigations provide a basis for constructing a topology of color experience that is more accurate than a linear scale or simple double-opponent color wheel (MacLennan, 1999b). In some cases we can predict the phenomenology of neurologically abnormal vision, and similar approaches allow us to at least begin to construct the perceptual experiences of non-human animals.

We have seen that the plausibility of spectral inversions depends on superficial phenomenological analysis, whereas more careful neurophenomenological investigation begins to reveal the inevitable structure of perceptual experience. Therefore, it is worth recalling that even apparently simple phenomena, such as color or pitch, have connections to other aspects of our experience. In fact, it is a mistake to assume that color terms refer primarily to wavelengths of light. For example, translators have been perplexed by the ancient Greek word *chlôros* (which nominally means green) because ancient texts apply it to blood, dew, tears, and other things that are not green in color (Gage, 1993, p. 272n7; Zajonc, 1993, p. 15). The explanation is that in ancient Greek, as in English, things that are moist, green, or living can be described as "green"; for example, we can speak of a green twig or a green rider without meaning they are green in color. Similarly, many other color terms were originally monovalent terms for minerals, dyes, and other substances, and seem to be polyvalent only when supposed to refer to ranges of wavelengths; for example, Medieval scarlets may be green, blue, black, or white in color (Gage, 1993, pp. 34–5). Therefore; we must expect that a comprehensive phenomenology of color (and other perceptual qualities) will include an extensive penumbra of material, emotional, and other associations, both phylogenetic and ontogenetic.

The Unity of Consciousness

The protophenomenal approach can provide some insights into the question of the unity of consciousness. For just as there is no reason to postulate a reified phenomenon to integrate the coherent activities of protophenomena into a whole, so there is no reason to postulate a separately existing subject to integrate the totality of phenomena into a unified conscious experience. Rather, the unity of consciousness consists in the dense network of interdependencies among the protophenomena, which is the causal nexus of the phenomenal world.

This conclusion is supported by empirical evidence from cerebral commissurotomies (split-brain operations), which sever the corpus callosum, the thick band of 800 million nerve fibers that connects the cerebral hemispheres (e.g., Gregory, 1987, pp. 740–7). The effect is to separate one consciousness (one subject) into two, which is just what we would expect, since the surgery eliminates connections between activity sites, and thus removes dependencies between protophenomena.

Indeed, we may suppose that as the nerve fibers are severed the protophenomena associated with the two hemispheres are progressively decoupled; so the one phenomenal world gradually divides into two, which implies that the unity of consciousness is a matter of degree. I don't know if the experimental evidence is available, but the claim certainly has empirical content.

One kind of evidence results from fact that these operations leave the brain stem intact. Thus some connections between the hemispheres remain, pro-

ducing two loosely coupled phenomenal worlds, whereas we would expect a complete bilateral section of the brain to produce two completely independent phenomenal worlds. Interestingly, it has been observed that the two hemispheres of these patients may communicate with each other by means of "external transactions", such as twitching the skin of the face, a process of which both subjects are, apparently, unconscious. Nevertheless, these transactions establish loose dependencies between the two phenomenal worlds (in one patient's head), which differ only in degree from the dependencies established when two people (each their own phenomenal world) interact. Mirror neurons, which mimic the activity of neurons in another person's brain (Rizzolatti & Craighero, 2004), also suggest that different individuals' phenomenal worlds may be more closely connected than we have been accustomed to think. These considerations suggest that the unity of consciousness should be placed on a continuum that includes progressively more loosely coupled phenomenal worlds.

The Unconscious Mind

I have hypothesized, on the basis of parsimony, that all neurons have activity sites, and therefore that all neurons have associated protophenomena, but this would seem to leave no place for the unconscious mind. For this discussion, the exact sense of "the unconscious mind" is not critical, and Jung's definition will suffice:

> "Everything of which I know, but of which I am not at the moment thinking; everything of which I was once conscious but have now forgotten; everything perceived by my senses, but not noted by my conscious mind; everything which, involuntarily and without paying attention to it, I feel, think, remember, want, and do; all future things that are taking shape in me and will sometime come into consciousness: all this is the content of the unconscious." (Jung, 1960, § 382)

However, there are at least three ways in which protophenomenal theory can accommodate the unconscious mind.

First, recall that protophenomena are not phenomena; although protophenomena bear elementary subjectivity, typically they are not individually salient in consciousness. Only by coherent activity do protophenomena emerge as distinct phenomena in the conscious state. Conversely, incoherently active protophenomena form a sort of background noise in the conscious state. We may compare the motion of air molecules, which is salient only if coherent (wind or a breeze), but is unnoticed if it is random.

The brainstem, midbrain, and right cerebral hemisphere have been mentioned as likely substrates for the unconscious mind (e.g., Stevens, 2003, ch. 13), but there is no reason to suppose that neurons in these areas do not have protophenomena, whereas those in the (manifestly conscious) left cerebral hemisphere do. Experiments with split-brain patients suggest a resolution of this paradox, for their brains house two conscious minds, each unaware of the other, that is, each an unconscious mind from the perspective of the other. As previously discussed, the hemispheres are capable of limited communication by means of the intact brainstem and "external transactions", but these communications from one hemisphere are experienced by the other hemispheric consciousness as inexplicable "hunches", just like those from the unconscious (Gregory, 1987, p. 743).

Therefore we may hypothesize that the normal brain houses several loosely communicating consciousnesses, that is, several loosely coupled phenomenal worlds, each a consciousness in itself, but experiencing the others as uncon-

scious minds.[18] This would accord with Jung's (1960, § 253) observation that unconscious complexes and archetypes often behave as autonomous personalities, who interact with ego consciousness by means of hunches, intuitions, compulsions, resistances, moods, dreams, and a variety of neuroses.

It might seem unlikely that these semi-independent conscious minds could exist unseen in the normal brain, but we must recall that initially it was not obvious that the split-brain patients had two conscious minds; they appeared perfectly normal until laboratory testing revealed anomalies. Similarly, in the normal human it may be the mind that includes the verbal and motor protophenomena that is most able to manifest its existence in behavior and is most easily identified with the ego. Other minds, which are more remote from the verbal and motor protophenomena (in terms of protophenomenal control), are less able to manifest their existence in observable behavior; they normally escape notice. So the second protophenomenal explanation of the unconscious mind is that it is not unconscious in itself, but only from the perspective of ego consciousness.

The third explanation is based on the hypothesis of Pribram and Sherrington, discussed above, that conscious experience is associated with graded electrochemical processes in the dendrites, but not with the all-or-nothing generation of action potentials in the axons.[19] That is, the activity sites reside in the dendrites, but not in the axons. This hypothesis accords well with Jung's account of "the archetypes of the collective unconscious", which he described as contentless perceptual-behavioral patterns grounded in our biological (and even physical) nature:

> "Again and again I encounter the mistaken notion that an archetype is determined in regard to its content, in other words, that it is a kind of unconscious idea (if such an expression be admissible). It is necessary to point out once more that archetypes are not determined as regards their content, but only as regards their form and then only to a very limited degree. A primordial image is determined as to its content only when it has become conscious and is therefore filled out with the material of conscious experience. … The archetype in itself is empty and purely formal, nothing but a *facultas praeformandi*, a possibility of representation which is given *a priori*. The representations themselves are not inherited, only the forms, and in that respect they correspond in every way to the instincts, which are also determined in form only." (Jung, 1969a, § 155)

That is, the archetypes reside in the axonal structures (nerve fibers), which are for the most part genetically and developmentally determined. According to Jung, when an archetype is activated and emerges into consciousness, it does so with specific phenomenal content. This conscious content is (according to the Sherrington-Pribram hypothesis) a consequence of graded interactions in the dendrites, the structure of which is largely a function of individual (vs. phylogenetic) development, learning, and adaptation.

Thus we have three different explanations of the unconscious mind, each compatible with protophenomenal theory and with each other. They may operate individually or in combination to produce the unconscious mind (a negative concept, defined by the *absence* of external evidence of consciousness).

Degrees of Consciousness

According to protophenomenal theory, protophenomena are associated with activity sites in the brain, and the structure of a phenomenal world corresponds to the interconnections among the activity sites. Fewer activity sites imply fewer degrees of freedom in a phenomenal world. Therefore, we would expect animals with simpler nervous systems than ours to have correspondingly simpler phenomenal worlds (fewer degrees of freedom, simpler structure).[20] However, there are a number of issues that cannot be resolved without

neurophenomenological investigation. For example, some animals (such as whales and elephants) have larger brains than we do (in part, to accommodate their larger bodies), and so we expect that their phenomenal worlds have more degrees of freedom, but it is not simply a matter of numbers, for phenomenal worlds (that is, consciousnesses) can be radically different in structure. Another issue, of course, is the nature of the activity sites, since if they are restricted to particular kinds of neurons, synapses (e.g., chemical vs. electrical), or neurotransmitters, then some other kinds of animals may have many fewer protophenomena than we do. The question has empirical content, but is very difficult to answer at this time.

Alternative Activity Site Hypotheses

We have used as a working hypothesis that the activity sites are in the dendritic trees of neurons. In contrast, Cook (2000, 2002a, 2002b, chs. 6–7) has suggested that the axon hillocks, where action potentials are initiated, are the activity sites and that the presence of a protophenomenon in conscious experience corresponds to the opening of several hundred thousand ion channels when the neuron fires; under these circumstances the intra- and extracellular fluids are not separated, and the cell, in effect, senses its (cellular) environment; the distinction between "self" and "other" is momentarily dissolved:

> "… the momentary opening of the cell membrane at the time of the action potential is the single-cell protophenomenon… underlying 'subjectivity' – literally, the opening up of the cell to the surrounding biochemical solution and a brief, controlled breakdown of the barrier between cellular 'self' and the external world" (Cook, 2002a).

Synchronous neural firing corresponds to the coherence of protophenomena into phenomena, and so

> "… the normal ebb-and-flow in the strength of subjective feeling is real, and a direct consequence of the variable number of neurons participating in synchronous firing" (Cook, 2002a).

According to Cook's theory, while neural firing is the physical correlate of *consciousness* (experience), physical processes in the dendrites are the correlates of *cognition* (information processing).

Others, more controversially, have suggested that consciousness is associated with the brain's electromagnetic (EM) field (John, 2002; McFadin, 2002, 2007; Pockett, 2000, 2002, 2007), and evidence has been adduced that it can affect neuron firing (McFadden, 2002). More specifically, McFadden hypothesizes (1) that neural firing induces an *endogenous EM field*, that this field influences neural activity, and that this feedback through the endogenous EM field is essential to neural information processing, and (2) reportable conscious experience (i.e., conscious experience that can result in publicly observable behavior) is associated with a component of this field that affects motor neurons.[21]

18
This hypothesis does not exclude the first possibility, namely that protophenomena in some areas, such as the brainstem and midbrain, are not sufficiently coherent to constitute phenomena.

19
See Miller, Galanter & Pribram (1960, pp. 23–4) and Pribram (1971, pp. 104–5, 1991, pp. 7–8).

20
Chalmers (1995) reaches the same conclusion on the basis of his "double aspect principle".

21
This is similar to the issue of non-reportable conscious experience discussed in connection with the unconscious mind.

If McFadden's hypotheses are correct, then there are several interesting implications for the theory of protophenomena. The first hypothesis implies that the EM field can mediate interactions among activity sites, and therefore that field effects might be relevant to protophenomenal interdependencies, which could be more diffuse and holistic than those corresponding to neural structures (see below on protophenomenal dependencies). The second hypothesis raises the possibility that some activity sites may be located in the endogenous EM field.

This possibility is reinforced by Dennis Gabor's (1946) analysis of the information carrying capacity of arbitrary signals (reviewed in MacLennan, 1991). He applied the Heisenberg-Weyl derivation of the Uncertainty Principle to prove a minimum joint localization in any two conjugate variables (e.g., time and frequency), and therefore that any finite, band-limited signal has a maximum number of degrees of freedom that may be used to convey information, its *logon content*.[22] This maximum is achieved by decomposing the signal into a superposition of *Gabor wavelets* (Gaussian-modulated complex exponentials, equivalent to the *pure states* of quantum mechanics), which are in effect quanta of information (called *logons*). Information is represented in the (complex-valued) coefficients of the logons. As a consequence the physical activity sites are localized but distributed patches of the EM field of various spatial frequencies with various orientations; they may be visualized as oriented grating patches. Activity is represented in the amplitude and phase of each patch, which raises the question of how the amplitude and phase of the protophenomena could differently affect conscious experience.

To determine the logon content of the brain's endogenous EM field, the relevant conjugate variables are area and spatial frequency. McFadden (2002) states that the spatial resolution of the field is smaller than 1 mm. From a cortical area of 2200 cm^2 we can calculate an approximate logon content of 2200 cm^2 / $(0.1\ cm)^2$ = 220 000 logons.[23] If the resolution were as fine as 0.1 mm (which is still quite coarse in neural terms; microcolumns have diameters an order of magnitude smaller: Jones, 2000), then the field could support approximately 22 million logons. Therefore, if the logons of the brain's endogenous EM field are activity sites, then each of our phenomenal worlds comprises some hundreds of thousand or millions of protophenomena (the intensities of which are correlated to the corresponding Gabor coefficients). Of course the existence of activity sites in the EM field does not contradict their existence in neurons as well. These issues can be addressed empirically, but I do not think we have the technology yet.

Nonbiological Consciousness

What Physical Processes Have Protophenomena?

I have discussed protophenomena in terms of human consciousness, but it is now time to consider them also in the context of robot consciousness. The crucial question is whether robot brains can be made sufficiently similar to human brains *in the relevant ways*. This can be explained by analogy. Liquidity is a property of water, but it depends on more fundamental physical properties of H_2O molecules, such as their finite volume and mutual attraction at close distances. Therefore, other substances that have these same fundamental properties, but are otherwise dissimilar to water, may be liquid. Similarly, protophenomena are a consequence of certain (currently unknown) physical

properties of activity sites. They might be quite specific to neurons, or they might occur in other physical systems as well. In the latter case, it would be reasonable to suppose that nonbiological systems with these properties would have artificial activity sites and corresponding protophenomena. If the activity sites were appropriately structured (also very poorly understood), then the protophenomena would cohere into phenomena and constitute a conscious state.

Obviously, the question cannot be answered without adequate knowledge of the activity sites associated with the protophenomena of human consciousness, but I can outline some of the possibilities.

Suppose protophenomena are associated with neural somata and that protophenomenal intensity corresponds to the membrane potential. If the robot's brain is not made from biological neurons, then the question becomes whether the biological character of the neuron is a necessary condition for it to have an associated protophenomenon. If, on the other hand, the presence of a protophenomenon depends only on certain electrochemical processes occurring in the cell body, it might be possible to construct an artificial device implementing those electrochemical processes and therefore having an associated protophenomenon. (By the way, it is difficult, though not impossible, to answer this question empirically, for phenomenological observation can establish the presence or absence of coherent ensembles of protophenomena, and perhaps in some cases of isolated protophenomena.)

Suppose instead that protophenomena are associated with synapses and their intensity with neurotransmitter flux. This raises a further question (which can be answered empirically): are protophenomena associated with all neurotransmitters and their receptors, or only with certain ones? If only with certain ones, then we have the further empirical question of why certain neurotransmitters should be associated with protophenomena but not others. What is the relevant difference between the neurotransmitters or between their receptors? When we know the answer to this question, then we can say whether the constituents of a robot's brain have the relevant properties to have protophenomena.

If, on the other hand, as Cook suggests, protophenomenal intensity corresponds to the opening of the cell to its environment and ion flux through the membrane, then we will need to discover whether any such boundary opening suffices for protophenomenal intensity, or only in the context of a living cell maintaining its existence as an entity distinct from its environment.

Similarly, if McFadden is correct in his connection of the brain's electromagnetic field with conscious experience, then to answer the question for robots we will need to understand what aspects of the mutual coupling of neurons and their EM field are relevant to conscious experience.

In summary, although these questions are complex and difficult, they are not unanswerable. The experiments are challenging, but not impossible.

A very interesting possibility is raised by Chalmers (1996, ch. 8). We have seen that protophenomena are essentially quality-less and that they acquire their qualities only through their mutual interdependencies; that is, the subjec-

22
Gabor's theory treats *structural information*, whereas Shannon's better-known theory treats *selective information*; they are complementary (see Cherry, 1978, pp. 47–49; MacKay, 1969, pp. 178–189; MacLennan, 1991).

23
The exact value depends on how spatial resolution is measured (see MacLennan, 1991), but the order of magnitude is correct.

tive quality is structured by formal relations among abstract quantities (protophenomenal intensities). (Although abstract, they are experienced, for the intensity of a protophenomenon is the degree of its presence in conscious experience.) Consistently with this, Chalmers suggests that *physically realized information spaces* might provide the link between the phenomenological and physical domains. When such a system is observed from the outside, we may give a physical account of its behavior, but when it is experienced from the inside, that is, when *I* am the physical information system, then I may have a subjective experience of the information processes. In other words, physically realized information spaces may be experienced objectively from the outside or subjectively from the inside.

Applied to protophenomena, this theory implies that any physically realized information space might be an activity site with an associated protophenomenon. Therefore, if the constituents of a robot's brain implement physically realized information spaces, as they surely must, then they would have associated protophenomena. This does not, in itself, imply that the robot will have conscious experience, for the protophenomena must be interdependent in such as way as to cohere into phenomena (i.e., conscious content), but if the robot's brain were structured to implement the functions of consciousness discussed in Section 2, then conscious experience would seem to be inevitable.

If Chalmer's idea is correct, then we must ask what constitutes something as a physically realized information space. We have Shannon's and Gabor's complementary information theories, which allow us to quantify information and changes in information state. For example, we can quantify the information received when an ion channel opens or a ligand binds to a receptor on a cell membrane, information that is used in governing later cellular processes (MacLennan, in press). It is plausible that these channels and receptors are activity sites, and that the ion flux or receptor activation corresponds to protophenomenal intensity. If this is true, then protophenomena need not be confined to neurons or even to eukaryotic cells.

If we take a further step, and accept Wheeler's (1994) ontological maxim, "it from bit", which asserts that all physical processes are fundamentally information processes, then we must entertain the possibility that all fundamental physical processes (such as quantum state change, or objective wave function collapse) have associated protophenomena.[24] This does not imply that computers, the earth, or the entire universe are conscious, for that would require that the protophenomena act in a sufficiently coherent and structured manner to constitute phenomena. This would be *panpsychism*, a much stronger claim than *panprotophenomenalism*, which asserts only that elementary subjectivity accompanies physical processes (a strong enough claim already, to be sure!). Panprotophenomenalism does not imply ubiquitous consciousness.[25] Interesting though these speculations may be, at this time we need to focus our investigations on the only protophenomena that we know exist: those associated with human brains.

Why Should We Care?

It may be worthwhile to make a few remarks about why we should be concerned about the Hard Problem for robots. If the robot does its job effectively, why should we care whether it is aware that it is doing it? One (perhaps distant) reason is the issue of robot rights. We do not have to go so far as imagining androids with human-like behavior, because the problem may arise with simpler machines, for rights are frequently grounded (often implicitly) in the

capacity to suffer. Cruel practices, such as vivisection, have been justified by the claim that "beasts" (non-human animals) are "just machines", a view that became widespread with the ascendancy of the mechanical philosophy of Gassendi and Descartes. (According to this philosophy, humans – or at least some humans! – were considered more than machines because they have "immortal souls"; in contrast, animals were considered soulless.) Nowadays, although there is ongoing debate about the existence and extent of animal rights, we do acknowledge animal suffering and try to avoid it (at least for some animals: cattle, but chickens? lobsters? oysters?).[26] So I think it is likely that we will face similar issues regarding sophisticated autonomous robots (especially those made out of organic materials).

A more immediate reason for worrying about the Hard Problem for robots is that it is a valuable test case for our understanding of our own conscious selves. If we cannot give a principled explanation why robots can or cannot have subjective experiences, then we do not understand our own consciousness very well. So long as we cannot answer the question for robots, the explanatory gap between mind and matter remains.

4. Conclusions

The "less hard" problems of consciousness relate to its functions in perception, cognition, and behavior, which in the case of animals can be determined by reference to the selective advantage of these functions in the species' environment of evolutionary adaptedness. Since these functions are also valuable for autonomous robots, I anticipate that robots will have to implement these functions as well, which will require solving the "less hard" (but nevertheless very difficult!) problems of functional consciousness and its physical mechanisms.

Closely related to consciousness is the issue of intentionality, the "aboutness" of functionally conscious (and other) brain states. I argued that intrinsic intentionality is grounded in the relevance of an agent's representations to the continued existence of the agent or its group, and so intentionality is largely independent of consciousness; indeed, very simple agents (organisms and machines) can exhibit genuine intrinsic intentionality. Nevertheless, truly autonomous robots must take care for the survival of themselves and others, and so intrinsic intentionality will characterize many of their internal states, including functionally conscious states.

Finally, I turned to the Hard Problem – how we can reconcile physical mechanism with the experience of subjective awareness – and addressed it from the perspective of neurophenomenology and the theory of protophenomena. Unfortunately, the possibility of a (sufficiently complex) robot having subjective experience cannot be answered without a better understanding of the relation of protophenomena to their physical activity sites. I considered several possibilities discussed in the literature and their implications for robot consciousness. Perhaps the most intriguing and parsimonious possibility is that

24
In this connection it is interesting to recall that Gabor's quantum of information, the *logon*, is mathematically identical to a *pure state* in quantum mechanics, and obeys the same Uncertainty Principle.

25
Panprotophenomenalism is of course a variety of double-aspect monism, discussed previously (footnote 13).

26
On these issues, see especially Beckoff (2007).

protophenomena are the "interior" aspects of physically realized information spaces. If this were so, then it would be highly likely that autonomous robots possessing functional consciousness with intrinsic intentionality would also experience subjective awareness. In such robots, there would be somebody home.

References

Adelman, G. (ed.) (1987) *Encyclopedia of Neuroscience*, Boston: Birkhäuser.

Bavelier, D., and Neville, H. J. (2002) "Cross-modal plasticity: Where and how?" *Nature Reviews Neuroscience*, vol. 3, pp. 445–452.

Bekoff, M. (2007) *The Emotional Lives of Animals. A Leading Scientist Explores Animal Joy, Sorrow, and Empathy – and Why They Matter*, New York: New World Library.

Bendor, D., and Wang, X. (2005) "The neuronal representation of pitch in primate auditory cortex", *Nature*, vol. 436, pp. 1161–5.

Berlin, B., and Kay, P. (1969) *Basic Color Terms. Their Universality and Evolution*, Berkeley: University of California.

Blackburn, S. (1994) *The Oxford Dictionary of Philosophy*, Oxford: Oxford University Press.

Block, N. (1995) "On a confusion about a function of consciousness", *Behavioral and Brain Sciences*, vol. 18, pp. 265–66.

Brentano, F. (1995) *Psychology from and Empirical Standpoint*, trans. A. C. Rancurello, D. B. Terrell, and L. L. McAlister, London & New York: Routledge.

Bridgeman, B. (1982) "Multiplexing in single cells of the alert monkey's visual cortex during brightness discrimination", *Neuropsychologia*, vol. 20, pp. 33–42.

Brooks, R. A. (1987) "A Hardware Retargetable Distributed Layered Architecture for Mobile Robot Control", in *Proceedings IEEE Robotics and Automation*, Raleigh, NC, pp. 106–110.

Burghardt, G. M. (1970) "Defining 'Communication'", in J. W. Johnston, Jr., D. G. Moulton, and A. Turk (eds.), *Communication by Chemical Signals*, New York: Appleton-Century-Crofts, pp. 5–18.

Buss, D. M. (2004) *Evolutionary Psychology. The New Science of the Mind*, 2nd ed., Boston: Pearson.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., Woodruff, P.W., Iversen, S. D., and David, A. S. (1997) "Activation of auditory cortex during silent lipreading", *Science*, vol. 276, pp. 593–596.

Cangelosi, A., and Parisi, D. (eds.) (2001) *Simulating the Evolution of Language*, London: Springer.

Chalmers, D. J. (1995) "Facing up to the problem of consciousness", *Journal of Consciousness Studies*, vol. 2, pp. 200–219.

Chalmers, D. J. (1996) *The Conscious Mind*, New York: Oxford University Press.

Chalmers, D. J. (2002) "Consciousness and its place in nature", in *Philosophy of Mind: Classical and Contemporary Readings*, D. Chalmers (ed.), Oxford: Oxford.

Cherry, C. (1978) *On Human Communication*, Cambridge, MA: MIT.

Cook, N. D. (2000) "On defining awareness and consciousness: The importance of the neuronal membrane", in *Proceeding of the Tokyo-99 Conference on Consciousness*, Singapore: World Scientific.

Cook, N. D. (2002a) "Bihemispheric language: How the two hemispheres collaborate in the processing of language", in *The Speciation of Modern Homo Sapiens*, T. Crow (ed.), London: Proceedings of the British Academy.

Cook, N. D. (2002b) *Tone of Voice and Mind. The Connections Between Intonation, Emotion, Cognition and Consciousness*, Amsterdam: John Benjamins.

Dennett, D. C. (1987) *The Intentional Stance*, Cambridge, MA: MIT Press.

De Valois, R. L., and De Valois, K. K. (1988) *Spatial Vision*, New York: Oxford.

De Valois, R. L., and De Valois, K. K. (1993) "A multi-stage color model", *Color Vision*, vol. 33, 1053–65.

Dretske, F. (1985) "Machines and the mental", *Proceedings and Addresses of the American Philosophical Association*, vol. 59, pp. 23–33.

Dreyfus, H. L. (1991) *Being-in-the-World. A Commentary on Heidegger's* Being and Time*, Division I*, Cambridge, MA: MIT Press.

Dunny, G. M., and Winans, S. C. (eds.) (1999) *Cell-cell Signaling in Bacteria*, Washington, D.C.: ASM Press.

Gabor, D. (1946) "Theory of communication", *Journal of the Institution of Electrical Engineers*, vol. 93, pt. III, pp. 429–57.

Gage, J. (1993) *Color and Culture. Practice and Meaning from Antiquity to Abstraction*, Boston: Little, Brown & Co.

Gaulin, S. J. C., and McBurney, D. H. (2004) *Evolutionary Psychology*, 2nd ed., Upper Saddle River: Pearson.

Goethe, J. W. von (1840) *Goethe's Theory of Colours*, trans. C.L. Eastlake, London: Murray.

Gould, E., Reeves, A., Graziano, M., and Gross, C. (1999) "Neurogenesis in the neocortex of adult primates", *Science*, vol. 286, pp. 548–52.

Gregory, R. L. (ed.) (1987) *The Oxford Companion to the Mind*, Oxford: Oxford University Press.

Grice, H. P. (1957) "Meaning", *Philosophical Review*, vol. 66, pp. 377–88.

Gutenplan, S. (ed.) (1994) *A Companion to the Philosophy of the Mind*, Oxford: Blackwell.

Hardin, C.L. (1988) *Color for Philosophers. Unweaving the Rainbow*, Indianapolis & Cambridge: Hackett.

Haugeland, J. (ed.) (1997) *Mind Design II. Philosophy, Psychology, Artificial Intelligence*, Cambridge, MA: MIT Press.

Heidegger, M. (1962) *Being and Time*, trans. J. Macquarrie and E. Robinson, New York: Harper & Row.

Heidegger, M. (1982) *The Basic Problems of Phenomenology*, trans. A. Hofstadter, Bloomington: Indiana University Press.

Hempel, C. G. (1965) *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, New York: The Free Press.

Hering, E. (1878) *Zur Lehre vom Lichtsinne*, Wien: Carl Gerold's Sohn.

Husserl, E. (1931) *Ideas. General Introduction to Pure Phenomenology*, trans. W. R. B. Gibson, New York: Macmillan.

Husserl, E. (1960) *Cartesian Meditations. An Introduction to Phenomenology*, trans. D. Cairns, The Hague: Nijhoff.

Husserl, E. (1973) *Experience and Judgment. Investigations in a Genealogy of Logic*, trans. J. S. Churchill and K. Ameriks, Evanston: Northwestern University Press.

Ihde, D. (1986) *Experimental Phenomenology. An Introduction*, Albany: State University of New York Press.

John, E. R. (2002) "The neurophysics of consciousness", *Brain Research Reviews*, vol. 39, pp. 1–28.

Jones, E. G. (2000) "Microcolumns in the cerebral cortex", *Proceedings of the National Academy of Sciences USA*, vol. 97, pp. 5019–21.

Jung, C. G. (1960) *The Structure and Dynamics of the Psyche* (*Collected Works*, vol. 8), 2nd ed., trans. R. F. C. Hull, Princeton: Princeton University Press.

Jung, C. G. (1969a) *The Archetypes and the Collective Unconscious* (*Collected Works*, vol. 9, part i), 2nd ed., trans. R. F. C. Hull, Princeton: Princeton University Press.

Jung, C. G. (1969b) *Psychology and Religion: West and East* (Collected Works, vol. 11), 2nd ed., trans. R. F. C. Hull, Princeton: Princeton University Press.

Jung, C. G., and Pauli, W. (1955) *The Interpretation of Nature and the Psyche*, trans. R. F. C. Hull and P. Silz, London: Routledge & Kegan Paul.

Kaiser, P. K., and Boynton, R. M. (1996) *Human Color Vision*, 2nd ed., Washington: Optical Society of America.

Karl, A., Birbaumer, N., Lutzenberger, W., Cohen, L. G., and Flor, H. (2001) "Reorganization of motor and somatosensory cortex in upper extremity amputees with phantom limb pain", *The Journal of Neuroscience*, vol. 21, pp. 3609–18.

Kay, P., and McDaniel, C.K. (1978) "The linguistic significance of the meanings of basic color terms", *Language*, vol. 54, 610–46.

Knudsen, E. J., du Lac, S., Esterly, S. D. (1987) "Computational maps in the brain", *Annual Review of Neuroscience*, vol. 10, pp. 41–65.

Laughlin, C. D., Jr., McManus, J., and d'Aquili, E. G. (1990) *Brain, Symbol and Experience. Toward a Neurophenomenology of Consciousness*, Boston: New Science Library.

MacKay, D.M. (1969) *Information, Mechanism and Meaning*, Cambridge, MA: MIT.

MacLennan, B. J. (1990) *Evolution of Communication in a Population of Simple Machines* (Technical Report UT-CS-90–99), Knoxville: University of Tennessee, Knoxville, Department of Computer Science.

MacLennan, B. J. (1991) *Gabor Representations of Spatiotemporal Visual Images* (Technical Report UT-CS-91–144), Knoxville: University of Tennessee, Knoxville, Department of Computer Science.

MacLennan, B. J. (1992) "Synthetic ethology: An approach to the study of communication", in *Artificial Life II. The Second Workshop on the Synthesis and Simulation of Living Systems*, C. G. Langton, C. Taylor, J. D. Farmer, and S. Rasmussen (eds.), Redwood City: MIT Press, pp. 631–658.

MacLennan, B. J. (1995) "The investigation of consciousness through phenomenology and neuroscience", in *Scale in Conscious Experience. Is the Brain Too Important to be Left to Specialists to Study?*, J. King and K. H. Pribram (eds.), Hillsdale: Lawrence Erlbaum, pp. 25–43.

MacLennan, B. J. (1996a) "The elements of consciousness and their neurodynamical correlates", *Journal of Consciousness Studies*, vol. 3, pp. 409–424.

MacLennan, B. J. (1996b) *Protophenomena and their Neurodynamical Correlates* (Technical Report UT-CS-96–311), Knoxville: University of Tennessee, Knoxville, Department of Computer Science.

MacLennan, B.J. (1999a) "Neurophenomenological constraints and pushing back the subjectivity barrier", *Behavioral and Brain Sciences*, vol. 22, 961–3.

MacLennan, B. J. (1999b) *The Protophenomenal Structure of Consciousness with Especial Application to the Experience of Color: Extended Version* (Technical Report UT-CS-99-418), Knoxville: University of Tennessee, Knoxville, Department of Computer Science.

MacLennan, B. J. (2006) "Making meaning in computers: Synthetic ethology revisited", in *Artificial Cognition Systems*, A. Loula, R. Gudwin and J. Queiroz (eds.), Hershey: IGI Global, ch. 9 (pp. 252–83).

MacLennan, B. J. (in press) "Protophenomena: The elements of consciousness and their relation to the brain", *Irreducibly Conscious. Selected Papers on Consciousness*, A. Batthyány, A. Elitzur, and D. Constant (eds.), Heidelberg & New York: Universitätsverlag Winter, ch. X (pp. 189–214).

MacLennan, B. J., and G. M. Burghardt (1993) "Synthetic ethology and the evolution of cooperative communication", *Adaptive Behavior*, vol. 2, pp. 161–188.

Maxwell, G. (1980) "The ontological status of theoretical entities", in *Introductory Readings in the Philosophy of Science*, E. D. Klemke, R. Hollinger, and A. D. Kline (eds.), Buffalo: Prometheus Books.

McCall, R. J. (1983) *Phenomenological Psychology. An Introduction. With a Glossary of Some Key Heideggerian Terms*, Madison: University of Wisconsin Press.

McFaddin, J. (2002) "Synchronous firing and its influence on the brain's electromagnetic field: Evidence for an electromagnetic theory of consciousness", *Journal of Consciousness Studies*, vol. 9, pp. 23–50.

McFaddin, J. (2007) "Conscious electromagnetic field theory", *NeuroQuantology*, vol. 5, no. 3, pp. 262–70.

Miller, G. A., Galanter, E., and Pribram, K. H. (1960) *Plans and the Structure of Behavior*, New York: Adams-Bannister-Cox.

Nida-Rümelin, M. (1996) "Pseudonormal vision: An actual case of qualia inversion?", *Philosophical Studies*, vol. 82, 145–57.

Palmer, S. E. (1999) "Color, consciousness, and the isomorphism constraint", *Behavioral and Brain Sciences*, vol. 22, 923–89.

Pockett, S. (2000) *The Nature of Consciousness: A Hypothesis*, Lincoln: Iuniverse.

Pockett, S. (2002) "Difficulties with the electromagnetic theory of consciousness", *Journal of Consciousness Studies*, vol. 9, pp. 51–6.

Pockett, S. (2007) "Difficulties with the electromagnetic field theory of consciousness: An update", *NeuroQuantology*, vol. 5, no. 3, pp. 271–5.

Povinelli, D. J., and Cant, J. G. H. (1995) "Arboreal clambering and the evolution of self-conception", *The Quarterly Review of Biology*, vol. 70, pp. 393–421.

Pribram, K. H. (1971) *Languages of the Brain: Experimental Paradoxes and Principles in Neuropsychology*, Englewood Cliffs: Prentice-Hall.

Pribram, K. H. (1991) *Brain and Perception. Holonomy and Structure in Figural Processing*, Hillsdale: Lawrence Erlbaum.

Pribram, K. H., Spinelli, D.N., and Kamback, M.C. (1967) "Electrocortical correlates of stimulus response and reinforcement", *Science*, vol. 157, pp. 94–96.

Rakic, P. (2002) "Neurogenesis in adult primate neocortex: an evaluation of the evidence", *Nature Reviews Neuroscience*, vol. 3 (1), pp. 65–71.

Rizzolatti G., Craighero L. (2004) "The mirror-neuron system", *Annual Review of Neuroscience*, vol. 27, pp. 169–92.

Russell, B. (1927) *The Analysis of Matter*, London: Kegan Paul.

Saunders, B. A. C., and van Brakel, J. (1997) "Are there non-trivial constraints on colour categorization?", *Behavioral and Brain Sciences*, vol. 20, 167–228.

Searle, J. (1983) *Intentionality. An Essay in the Philosophy of Mind*, Cambridge, MA: Cambridge University Press.

Searle, J. (1992) *Rediscovery of the Mind*, Cambridge, MA: MIT Press, 1992.

Shear, J. (ed.) (1997) *Explaining Consciousness. The Hard Problem*, Cambridge, MA: MIT Press.

Shepherd, G. M. (1994) *Neurobiology*, 3rd ed., New York & Oxford: Oxford University Press.

Stevens, A. (2003) *Archetype Revisited. An Updated Natural History of the Self*, Toronto: Inner City Books.

Suga, N. (1984) "The extent to which biosonar information is represented in the bat auditory cortex", in G. M. Edelman, W. E. Gall, and W.M. Cowan (eds.), *Dynamic Aspects of Neocortical Function*, New York: Wiley, pp. 315–73.

Suga, N. (1989) "Principles of auditory information-processing derived from neuroethology", *Journal of Experimental Biology*, vol. 146, pp. 277–86.

Suga, N. (1994) "Processing of auditory information carried by species-specific complex sounds", in M. S. Gazzaniga (ed.), *The Cognitive Neurosciences*, Cambridge, MA: MIT, pp. 295–318.

Sur, M. (2004) "Rewiring cortex: Cross-modal plasticity and its implications for cortical development and function", in *Handbook of Multisensory Processing*, B. Stein (ed.), Cambridge, MA: MIT Press.

Wagner, K., Reggia, J. A., Uriagereka, J., and Wilkinson, G. S. (2003) "Progress in the simulation of emergent communication and language", *Adaptive Behavior*, vol. 11 (1), pp. 37–69.

Weiskrantz, L. (1995) "Blindsight: Conscious vs. unconscious aspects", in *Scale in Conscious Experience. Is the Brain Too Important to be Left to Specialists to Study?* J. King and K. H. Pribram (eds.), Hillsdale: Lawrence Erlbaum.

Wheeler, J. A. (1994) "It from bit", in *At Home in the Universe*, J. A. Wheeler, Woodbury: American Institute of Physics Press.

Zajonc, A. (1993) *Catching the Light. The Entwined History of Light and Mind*, New York: Bantam Books.

Bruce J. MacLennan

Natürliches und künstliches Bewusstsein

Zusammenfassung

Ausgehend von Erkenntnissen der Evolutionären Psychologie untersucht dieser Beitrag wichtige Funktionen, die das Bewusstsein autonomer Roboter ausfüllen kann. Gemeint sind willkürlich kontrolliertes Handeln, bewusstes Wahrnehmen, Eigenwahrnehmung, Metaerkenntnis sowie Bewusstsein des eigenen Selbst. Der Verfasser unterscheidet zwischen intrinsischer Intentionalität und Bewusstsein, führt jedoch das Argument ins Feld, dass es ebenso wichtig sei, die Erkenntnisweise eines Roboters zu verstehen. Abschließend wird, aus dem Blickwinkel der Theorie von den Protophänomenen, das für Roboter „schwierige Problem" untersucht, d.h. die Frage, ob sie zu subjektiver Wahrnehmung fähig sind.

Schlüsselbegriffe

Autonomer Roboter, Wahrnehmung (Gewahrsein), Bewusstsein, Evolutionäre Psychologie, das „schwierige Problem", Intentionalität, Metaerkenntnis, Protophänomene, Qualia, Synthetische Ethologie

Bruce J. MacLennan

La conscience, naturelle et artificielle

Résumé

En s'appuyant sur les résultats de la psychologie évolutionniste, nous examinons les différentes fonctions importantes que puisse remplir la conscience dans les robots autonomes : action contrôlée, prise de conscience, conscience de soi, métacognition, conscience du moi. Nous distinguons l'intentionnalité intrinsèque de la conscience, mais soutenons également l'importance de la compréhension de la cognition robotique. Enfin, nous étudions le « Hard Problem » concernant les robots, c'est-à-dire la question de savoir s'ils peuvent connaître une prise de conscience subjective, dans une perspective de la théorie du protophénomène.

Mots-clés

robot autonome, prise de conscience, conscience, psychologie évolutionniste, Hard Problem, intentionnalité, métacognition, protophénomène, qualia, éthologie synthétique