Consciousness in Robots:

The Hard Problem and Some Less Hard Problems
(Extended Version)

Technical Report UT-CS-05-553

Bruce J. MacLennan*

Department of Computer Science
University of Tennessee, Knoxville
www.cs.utk.edu/~mclennan

May 15, 2005

Abstract

Based on results from evolutionary psychology we discuss important functions that can be served by consciousness in autonomous robots. We distinguish intrinsic intentionality from consciousness, but argue it is also important. Finally we explore the Hard Problem for robots (i.e., whether they can experience subjective awareness) from the perspective of the theory of protophenomena.

* This report is an extended version of a paper for *14th IEEE International Workshop on Robot and Human Interactive Communication* (Aug. 13–15, 2005, Nashville, TN).

1 Introduction

There are many scientific and philosophical problems concerning consciousness, but in 1995 David Chalmers (1995) 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" (Chalmers 1995). The scientific investigation of experience is impeded by the unique epistemological status of consciousness (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 paper I will begin by considering some of these "less hard" problems in the context of robot consciousness, and then turn to the Hard Problem.

2 Less Hard Problems

2.1 The Functions of 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. In this section I will consider several of these notions and the "less hard" problems associated with them in the context of robotics.

What is consciousness good for? Is there any reason we should want our robots to be conscious? To answer these questions, we need to understand the *function*, 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 confers in the species' *environment of evolutionary adaptedness* (EEA), which is the environment in which it 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 this approach to address the "less hard" problems of robot consciousness. [An introduction can be found in many recent textbooks, such as Buss (2004) and Gaulin & McBurney (2004).]

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, eating with chopsticks for the first time, 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 animal stumbles over an unnoticed object. 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.

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 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 the use 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 through *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. 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 to the sound of a hostile agent) 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.

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, auditory, and olfactory interactions) can operate from a primarily subjective perspective, that is, the world is understood from an 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 environment, 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) will similarly require this kind of self-awareness in order to control their motion through a shared, objective environment. Therefore also they will 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"). 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. [This can be viewed as a specialized, high-level

application of Brooks' (1987) subsumption principle.] 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 (Stevens 2003), who observe that consciousness is not the central faculty of the psyche around which all the others orbit. 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 animals, including humans, that many of these functions may 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 conscious*, 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).

2.2 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. 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. (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.) 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 derived from 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, communication, etc.) (Dennett 1987).

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 understand 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 (for a fuller discussion see MacLennan 1992, MacLennan & Burghardt 1993). 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 the marks, may go in a different direction. Was this communication, or a non-communicative event in which the behavior of one animal indirectly influenced that of another? We would like to be able to determine if the *purpose* of the first animal's action was to influence the behavior of other animals, or if 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 as 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 representation 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 meaning).

Can artificial agents, such as robots, exhibit intrinsic intentionality? *Synthetic ethology* offers a methodology by which such questions can be addressed (MacLennan 1992, MacLennan & Burghardt 1993). The goal of synthetic ethology is to permit the scientific investigation of problems relating to the physical mechanisms 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. 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), 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 continuing "survival" (as organized structures) of the machines. 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.

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; 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 micro-organisms), then we must expect micro-evolution 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. (It might be mentioned in passing that emotions, which have many important connections to consciousness, are important in all these kinds of autonomous robotics.)

3 The Hard Problem

3.1 Why It Is Hard

Having discussed the "less hard" problems of robot consciousness (which are certainly hard enough to keep us busy for many years!), I will turn to the Hard Problem 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 (a fuller discussion can be found in Chalmers 1995, 1996, MacLennan 1995, 1996a, Searle 1992). The root of the problem is the unique epistemological status of consciousness, 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, science seeks 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.) 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 and Husserl. Consciousness is the vector of intentionality extending from the observer to the observed. 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. These same characteristics make conscious experience 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. Indeed, although third-person objects and properties can be reduced to other third-person objects and properties, it is a category mistake to attempt to reduce first-person phenomena to the third-person objects and properties.

3.2 Protophenomena

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).

The value of reductionism is that it allows us to understand higher-level phenomena better by relating them to lower-level phenomena. (Reductionism 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 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.) 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.

Indeed, neurologically-informed phenomenological reduction suggests that is may be fruitful to understand conscious experience in terms of *protophenomena*, which are theoretical entities hypothesized as the elementary constituents of phenomena (conscious experiences) (MacLennan 1995, 1996a).

The simplest kinds of protophenomena are similar to "sense data." 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. [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; see MacLennan (1991) for a survey.] However, 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 a philosophical context, a *phenomenon* is anything that appears in consciousness, and so 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 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).

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 vali-

dated by their fruitfulness for scientific inquiry (Hempel 1965, Maxwell 1980). 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 differ (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.) MacLennan (1995, 1996a) discusses the ontological status of protophenomena in more detail.

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 *dual-aspect monism*).

Unfortunately, I do not believe that we can say, at this time, what the activity sites are. Some reasonable possibilities include synapses and neural somata, in which cases the intensity of the associated protophenomenon might correspond to neurotransmitter flux or membrane potential. Cook (2000, 2002a, 2002b) has suggested that neurons are the activity sites and that the presence of a protophenomenon in conscious experience corresponds to the opening of several hundred 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. Others, more controversially, have suggested that consciousness is associated with the brain's electromagnetic field (John 2002, McFadin 2002, Pockett 2000, 2002), and evidence has been adduced that it can affect neuron firing (McFadden 2002). A simple analysis suggests that this field could represent a million information quanta (Gabor logons), and so an equal number of protophenomena (MacLennan 2003). In any case, the issue is unresolved, but in principle the question can be answered empirically, although I do not think we have the technology yet.

As implied in the foregoing description, 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). This intensity is hypothesized to be 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 experiences presence in consciousness. Alternately, if protophenomena correspond to the Gabor wavelets constituting the brain's electromagnetic field, then their intensities will correspond to the Gabor coefficients, which are complex numbers. This raises the question of how the amplitude and phase of the protophenomena will affect conscious experience. These are all scientific questions, which can be answered empirically.

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 an-

swer, 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, 2003). That is, the intensity of each protophenomenon is a complicated function of the recent intensities of (typically) thousands of other protophenomena, as well as of *extrinsic variables*, that is, of variables external to the phenomenological domain. From this perspective 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. Therefore, qualia are emergent properties in a phenomenological world structured by protophenomenal dependencies. (This is essentially a *structuralist* theory of qualia.)

3.3 Protophenomena and Robot Consciousness

I have discussed protophenomena in terms of human consciousness, but it is now time to return to the topic at hand, robot consciousness. The crucial question is whether robot brains can be made sufficiently similar to human brains *in the relevant ways*. Unfortunately, the question is difficult to answer without adequate knowledge of the activity sites associated with protophenomena, 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.)

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 protophenomena but not others? What is the relevant difference between the chemicals or 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 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 (1997). We have seen that protophenomena are essentially quality-less and that they acquire their qualities only through their mutual interdependencies; that is, the subjective 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 Part II, then conscious experience would seem to be inevitable.

3.4 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 soulless.) Nowadays, although there is ongoing debate about the existence and ex-

tent of animal rights, we do acknowledge animal suffering and try to avoid it (at least for some animals: cattle, but chickens? lobsters? oysters?). 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 their selective advantage 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 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

Brooks, R. A. (1987), "A hardware retargetable distributed layered architecture for mobile robot control," *Proceedings IEEE Robotics and Automation*, Raleigh NC, pp. 106–110.

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

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

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.

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, chs. 6–7.

Dennett, D. C. (1987), *The Intentional Stance*, Cambridge: MIT Press, ch. 8.

Gaulin, S. J. C., and D. H. McBurney (2004), *Evolutionary Psychology*, 2nd ed., Upper Saddle River: Pearson, ch. 5.

Hempel, C. G. (1965), *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, New York: The Free Press, pp. 177–9.

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

MacLennan, B. J. (1991), "Gabor representations of spatiotemporal visual images," technical report UT-CS-91-144, Dept. of Computer Science, University of Tennessee, Knoxville.

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, Dept. of Computer Science, University of Tennessee, Knoxville.

MacLennan, B. J. (1999), "The protophenomenal structure of consciousness with especial application to the experience of color: Extended version," technical report UT-CS-99-418, Dept. of Computer Science, University of Tennessee, Knoxville.

MacLennan, B. J. (2003), "Protophenomena: The elements of consciousness and their relation to the brain," technical report UT-CS-03-500, Dept. of Computer Science, University of Tennessee, Knoxville.

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, pp. 175–84.

J. McFaddin (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.

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.

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

Searle, J. (1992), *Rediscovery of the Mind*, Cambridge: MIT Press, 1992, chs. 4, 5.

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

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