A Protophenomenological Analysis of Synthetic Emotion in Robots

Technical Report UT-CS-08-623

Bruce J. MacLennan[*]

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

August 6, 2008

Abstract

This report addresses the "Hard Problem" of consciousness in the context of robot emotions. The Hard Problem, as defined by Chalmers, refers to the task of explaining the relation between conscious experience and the physical processes associated with it. For example, a robot can act afraid, but could it feel fear? Using protophenomenal analysis, which reduces conscious experience to its smallest units and investigates their physical correlates, we consider whether robots could feel their emotions, and the conditions under which they might do so. We find that the conclusion depends on unanswered but empirical questions in the neuropsychology of human consciousness. However, we do conclude that conscious emotional experience will require a robot to have a rich representation of its body and the physical state of its internal processes, which is important even in the absence of conscious experience.

[*] This report is an unedited draft of "Robot React, but Can They Feel?" (a chapter submitted for the *Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence*, edited by Jordi Vallverdú and David Casacuberta); it is being circulated for critical comment.

1 Introduction

Recent decades have seen a renaissance in the scientific investigation of consciousness, but a fundamental issue has been neglected, which is to integrate the facts of subjective experience with our understanding of physical processes in the nervous system. Chalmers (1995, 1996) has called this the *Hard Problem* of consciousness because it poses unique epistemological challenges that make it resistant to straight-forward scientific investigation (see also MacLennan, 1995, 1996). The fundamental problem is that it would seem possible for there to be "zombies" in which all the usual neurophysiological processes take place in the brain, leading to normal behavior, brain scans, etc., but without any accompanying subjective experience (Campbell, 1970; Kirk, 1974; Kripke, 1980).[1] Therefore, it is necessary to distinguish between *functional* (or *access*) *consciousness*, which refers to cognitive and behavioral functions fulfilled by consciousness in an organism, and *phenomenal consciousness,* which refers to the experience of subjective awareness (e.g., Block, 1995).

The Hard Problem is especially interesting when we consider robot emotions. Emotion is essential to the competent functioning of many animals (arguably, all animals: Panksepp, 2004, pp. 34–7; Plutchik, 2003, pp. 223–6), and synthetic emotion can fulfill similar functions in autonomous robots (as discussed briefly in **Section 2 Background**, below). Just as the emotion *fear* can galvanize an organism and reorganize its cognition and behavior in order to protect itself, so synthetic fear can function for a robot's self-protection. But will the robot feel afraid? Or, more carefully, what, if any, are the conditions under which a robot would *feel fear* (as opposed to *acting afraid*, in both its behavior and cognitive processes)? Thus, my goal in this report is to address *the Hard Problem of robot emotions.*

It bears repeating that the other problems of consciousness are not easy! If they are less hard, it is only because they are amenable to the usual methods of scientific investigation, and don't pose any unusual epistemological challenges. Certainly, functional consciousness is relevant to robot emotions, but this chapter will focus on phenomenal consciousness and robot emotions. (For a general discussion of robot consciousness, both functional and phenomenal, see MacLennan, 2008a.)

2 Background

2.1 The Biological Functions of Emotion

I begin by discussing the functions of emotion in the context of evolutionary biology and the relevance of these emotional functions to practical robotics; it is a brief overview, since the topic is addressed at length in other publications. But what is an emotion? Plutchik (2003, pp. 18–19) quotes some twenty definitions, and Paul and Anne Kleinginna (1981) develop a taxonomy of more than 90 definitions! For the purposes of this chapter, Rolls (2007) provides a good summary of the essential characteristics of emotion: An emotion is a state elicited by the delivery or omission of a reward or punisher (either

[1] Certain technical terms, such as "subjective," are defined in **Section 7 Key Terms and Definitions** (p. 25).

present or remembered), which functions as positive or negative reinforcement. Specifically, the delivery of a reward, or the omission or cessation of a punisher, is positive reinforcement, and conversely the delivery of a punisher, or omission or cessation of a reward, is negative reinforcement. Thus the organism acts to seek rewards and avoid punishers, and from an evolutionary perspective these actions are adaptive, in the sense of inclusive fitness (Plutchik, 2003, pp. 218–23). Rolls (2007) enumerates six principal factors in the elicitation of emotions: (1) reinforcement contingency (delivery, omission, cessation), (2) intensity, (3) reinforcement associations with a stimulus, (4) primary (i.e., phylogenetic) reinforcers of an emotion, (5) secondary (i.e., learned) reinforcers, and (6) the possibility of active or passive responses, which may affect the elicited emotion. (See also Rolls, 2002, 2005.)

What good is emotion? Why should we want robots to have them? Rolls (2005, 2007) lists nine significant functions (evolutionary adaptations) fulfilled by emotion:

(1) Emotion is essential in eliciting autonomic and endocrine responses, such as a change in heart rate and the release of adrenaline.

(2) Emotion facilitates flexibility of response, by separating the evaluation of a stimulus (as a reward or punisher) from the means to obtain or avoid it. That is, the emotion represents a *goal*, which might be achieved in a variety of ways (Rolls, 2006).

(3) The effect of emotion is inherently motivating (i.e., instigating action).

(4) Emotional expression facilitates communication by revealing an animal's mental state, goals, intentions, etc.

(5) Social bonding is also facilitated by emotion, for example between parents and offspring or among the members of a community.

(6) Emotion can persist over an extended period and generate a mood, which affects cognitive processing of events and memory, for example, biasing the processing to be more appropriate to the situation eliciting the emotion.

(7) Memory storage and retrieval are both facilitated by emotion, since emotionally charged events are more likely to be encoded into episodic memory; also, because the emotion is encoded and affects memory representation, similar emotions can aid the retrieval of memories that are relevant to an emotionally charged stimulus.

(8) Since an emotional state can persist over an extended period of time, it can provide a coherent continuing context for cognitive processing and behavior.

(9) Finally, emotion evokes retrieval of non-cortical memories.

2.2 Robot Emotions

Most of the functions enumerated by Rolls are also relevant to robotics; I will mention them briefly, citing the numbers of the functions in his list. Although a robot does not have an endocrine or autonomic nervous system, it may have analogous functions to accomplish, for example, redistributing power in a power-aware computer, powering-up inactive devices, reallocating computational resources, reorienting or tuning sensors, and readying effectors for action (1). Behavioral flexibility can be improved by decoupling the possible indicators of a situation from the possible responses to it, and this can be achieved by a "bow tie" organization in which the knot of the tie represents the motivational essence of a situation. (Bow tie organization is common in biological systems and is widely used in engineering for risk management.) The left-hand side of the bow tie represents various stimuli or signals that can indicate a general situation, while the right-hand side represents various responses to the situation. For example, many different stimuli can indicate to a robot that it is in a harmful environment, and that motivates the robot to plan some corrective action. In cases such as this, the "knot" is serving the function of an emotion, such as fear (2). A robot's activity is generally directed toward some broad goal (e.g., exploration, construction, refueling, defense, offense) which motivates and organizes its subsidiary activities, and which functions like an emotion (3); the persistence of an emotional state maintains this high-level organization until the goal is achieved or a higher priority goal arises (6, 8). Robots may have to work in collaborative teams including humans as well as other robots, and effective cooperation requires each agent to have some insight into the internal states, goals, and intentions of the other agents (Breazeal, 2003; Breazeal, Brooks, Gray, Hoffman, Kidd, Lee, Lieberman, Lockerd & Chilongo, 2004), for which overt display of emotional state is valuable (4, 5). Furthermore, it is important for robots that work with, assist, care for, or rescue humans to have feelings of care and concern for them, and conversely it may be useful for robots to feel a need for care (maintenance, repair, rescue, etc.) by humans or by other robots (5). Finally, the representation of emotion in memory and cognition provides a crude, but highly behavior-relevant cue for associative memory storage and retrieval and for context-sensitive cognitive processing (7, 9).

Though brief and superficial, the foregoing remarks are aimed at outlining the functional roles played by emotions in animals and at defending the usefulness of a similar structure of synthetic emotions for autonomous robots. Nevertheless, although modeled on natural emotion, there would seem to be no reason why robots could not have all the appropriate information structures and control processes to fulfill the functions of emotions, but without *feeling* them. That is, there would seem to be no contradiction in "zombie robots," who, for example, have an internal representation corresponding to fear, and react fearfully in appropriate circumstance, but which *feel* no fear. Regardless of whether one thinks this possibility is likely or unlikely, it remains in the realm of opinion unless we can find some principled, and preferably scientific, approach to the Hard Problem of robot emotions.

3 Protophenomenal Analysis of Robot Emotion

3.1 Overview of Protophenomenal Analysis

3.1.1 Neurophenomenology

Protophenomenal analysis aims to relate the structure of conscious experience to physical processes in the organism, and in particular it aims to understand the emergence of consciousness from its smallest constituents. Thus it takes subjective experience as an empirical given, and attempts to understand it in reference to neurobiology and underlying physical processes. Protophenomenal analysis does not presume that conscious experience can be reduced to physical processes (nor the converse!), but rather seeks mutually informative correspondences between the structures of the two domains (subjective and objective, as we might say).

Chalmers (1996) provides a good discussion of the Hard Problem and the background for protophenomenal analysis, although I do not agree with all of his conclusions. More detailed discussions of protophenomenal approach and its application to a number of problems in consciousness can be found elsewhere (MacLennan, 1995, 1996a, 1999b, 2008a, 2008b); here we restrict our attention to topics necessary for addressing emotion in robots.

In order to establish a detailed correspondence between conscious experience and neurophysiology, it is necessary to carefully investigate the structure of consciousness. As illustrated by the failings of naive introspectionism (Gregory, 1987, pp. 395–400; Lyons, 1986), this is a difficult undertaking. As in any empirical investigation, appropriate training in observation and experimental technique is required in order to see the relevant phenomena and make the appropriate discriminations. The empirical investigation of consciousness presents particular problems, due to the privateness of the phenomena and the effects that the observer can have on the phenomena, among other factors. Fortunately, work in phenomenological philosophy, including experimental phenomenology, and investigations in phenomenological psychology, have provided many empirical tools and techniques and produced an increasing body of results (e.g., Ihde, 1986; McCall, 1983).

A *phenomenon* may be defined as anything that appears or arises (Grk., *phainetai*) in conscious experience, including perceptions, but also hallucinations, ideas, recollections, expectations, hopes, fears, intentions, moods, feelings, and conscious emotions. A person's experience takes place in a *phenomenal world* comprising actual and potential phenomena experienceable by that person. In a broad sense *phenomenology* is the discipline that studies the structure of a phenomenal world, that is, the relationships of necessity, possibility, and probability among the appearances of phenomena, and especially those structures common to all people. To address the Hard Problem, however, we must resort to *neurophenomenology*, which seeks to correlate the dynamics of one's phenomenal world to the neurodynamics of the brain (Laughlin, McManus & d'Aquili, 1990; Lutz & Thompson, 2003; Rudrauf, Lutz, Cosmelli, Lachaux & Le Van Quyen, 2003; Varela, 1996). It combines the techniques of phenomenology and neuropsychology, using each to help us advance the other.

3.1.2 Neurophenomenological Reduction

Reduction is a valuable tool in science; it helps us to understand more complex systems in terms of simpler ones. However, the patterns of reduction that have been most successful in science are not applicable to the Hard Problem, for they reduce *objective* properties and processes to simpler objective properties, whereas the Hard Problem resides in the relation between the subjective and the objective, and a reduction of one to the other is fundamentally impossible (Chalmers, 1996, Pt. II; MacLennan, 1995, 1996).

Nevertheless, reduction is valuable, and so protophenomenal analysis makes use of reduction, but confined to either the subjective or objective domains. Phenomenological reduction, as applied in the subjective domain, seeks to relate subjective phenomena to simpler subjective phenomena.[2] Easiest to understand is a *qualitative reduction*, which divides up phenomena by modality; for example perceptual phenomena may be divided according to sensory modality. However careful phenomenology reveals that the various modalities are not completely independent, but interact in subtle and important ways, so most qualitative reductions are only approximate (MacLennan, 1999, 2003, 2008a, 2008b).

However, it is also possible to do a *quantitative reduction* in the subjective domain, in which subjective phenomena are reduced to simpler phenomena of the same kind. For example, visual phenomena are composed of smaller visual phenomena, such as subjective experiences of patches of color, oriented edges and textures, etc. Similarly, at least to a first approximation, tactile phenomena are constituted from much smaller phenomena corresponding to individual patches of skin. The reduction is more complicated, of course, when we consider such non-perceptual phenomena as intentions, beliefs, and — most relevant here — moods, emotions, etc., but with careful phenomenological analysis the reduction can be accomplished.

Phenomenological reduction can be paralleled by neurological reduction. For example, simpler phenomena (patches etc.) distributed across the visual field correspond to spatially distributed patterns of activity over visual regions of the cortex, called *retinotopic maps*. Similarly, *somatosensory maps* have a spatial structure corresponding systematically to the spatial organization of nerves in the skin, muscles, etc. Indeed, *topographic maps* are ubiquitous in the brain, and seem to be one of the fundamental ways the brain organizes the activity of individual neurons (or small groups of them, such as microcolumns) into macroscopic neural representations (Knudsen, du Lac & Esterly, 1987). Organization of the maps and corresponding phenomena need not be spatial. For example, *tonotopic maps* in auditory cortex systematically represent different pitches by the neurons in the map, corresponding to a reduction of subjectively perceived intervals of pitch into smaller subjective intervals.

In this way the phenomenological and neurological reductions facilitate each other: neurological discoveries suggest phenomenological structures, which may be explored through phenomenology, and phenomenological investigations provide observations to test and extend neurological theories. This joint reductive process may be termed *neurophenomenological reduction*, for it attempts a parallel reduction of subjective and objective processes to more elementary ones of the same kind. In this process neither domain is privileged over the other, for they are both empirically given, and both are necessary for solving the Hard Problem. At each stage of the process one or the other side may be experimentally or theoretically more tractable, and so each may facilitate progress in the other.

This parallel reduction has implications for the limit of neurophenomenological reduction, because apparently there are smallest units of neurological reduction (e.g., individual neurons or synapses, but it is not important at this point what they are). This implies that there will also be smallest units of phenomena, and we may call them *protophenome-*

[2] This notion of phenomenological reduction should not be confused with the phenomenological reduction described by Varela (1996), although they are not unrelated.

na (Chalmers, 1996, pp. 126–7, 298–9; Cook, 2000, 2002a, 2002b, chs. 6–7, 2008; MacLennan, 1996; cf. *proto-qualia* in Llinas, 1988; *phenomenisca* in MacLennan, 1995).

It is easy to confuse protophenomena with elementary sense data (e.g., in visual perception, patches of red-here-now, green-there-now, etc.), and some perceptual protophenomena are in fact similar to elementary sense data, but there are important differences. First, the neuropsychology of vision demonstrates that the elements of visual perception go far beyond patches of color, and include edges, lines, textures, and much higher level features. Further, protophenomena are extremely small, in comparison with ordinary phenomena, in the sense that an ordinary phenomenon comprises a very large number of protophenomena. This is implied by the neurophenomenological reduction, for if the smallest units of neural representation are individual neurons or microcolumns, then there could be many billions of protophenomena in the conscious state. Indeed, most protophenomena are unobservable, in the sense that a change of one protophenomenon will not usually be able to affect our behavior. The idea that the elements of consciousness are unobservable may seem paradoxical or even self-contradictory, but it is not, for *proto*phenomena are not phenomena. An analogy may clarify the situation: a diamond is a solid object made of carbon atoms, but a single carbon atom is not a diamond, and removing or adding a single atom will not affect the diamond *qua* diamond. (For more on the ontological status of protophenomena, see MacLennan, 1995, 1996a.) Finally, the neurophenomenological reduction implies that the protophenomena are the elementary constituents of *all* conscious phenomena, not just perceptions, and so we must use neurophenomenological investigations to describe the protophenomena of emotions and other nonperceptual phenomena.

It is also necessary to remark that phenomenology reveals that our ordinary conscious state is not passive experiencing of the world, but rather active, engaged *being-in-the-world* (Heideggarian *Dasein*: Dreyfus, 1991). Therefore, corresponding to sensory and motor neurons there are receptive and effective protophenomena, which are connected by more interior protophenomena corresponding to interneurons.

Sensory neurons usually have a *receptive field*, which defines their response to various stimuli. That is, the receptive field is a function that defines a neuron's degree of activity for different stimuli in its input space. For example, a neuron might show maximal activity for an edge of a certain orientation at a particular location on the retina, and its response might fall off for stimuli that are at different orientations, are less like edges, or are at slightly different locations. Interneurons also have receptive fields, but they are defined over more abstract input spaces than those of sensory neurons.

3.1.3 Activity Sites and Protophenomenal Intensity

From the phenomenological perspective, each protophenomenon has an *intensity* that measures its degree of presence in the conscious state. For example, corresponding to the above-mentioned neuron is a protophenomenon that contributes to phenomena an oriented edge at a particular location in the visual field. This protophenomenon's intensity constitutes its degree of presence in any phenomenon of which it is a part.

Neurophenomenological analysis implies that each protophenomenon has an associated *activity site* in the brain, and that some physical quantity or process at this site is necessarily correlated with the protophenomenon's intensity. (I say "necessarily correlated" to stress that protophenomenal intensity is not reducible to this physical quantity, or vice versa, but that their correlation is a necessary consequence of their being two aspects of the same reality; thus, the theory of protophenomena is a species of double-aspect monism (MacLennan, 1996a, 2008a), which Chalmers (2002) terms *type-F monism*.) Nevertheless we cannot say at this time what the activity sites might be. Obvious candidates include the neuron's somatic membrane, the synapses, and the dendritic trees or spines.

There are corresponding candidates for the physical process correlated with protophenomenal intensity, including membrane potential, ion or neurotransmitter flux, and the action potential. It is possible that different neurotransmitters lead to qualitatively different experiences, or that membrane depolarization is experienced differently from hyperpolarization, but in the absence of empirical evidence to the contrary, it is simpler to assume they are all experienced the same (i.e., degree of presence in consciousness). One attractive possibility, proposed by Cook (2000, 2002a, 2002b, chs. 6–7, 2008) is that the activity site is the neural membrane at the axon hillock (where the action potential is initiated), and that protophenomenal intensity corresponds to ion flux through the membrane when the neuron fires. In any case it is important to emphasize that the identities of the activity sites and of the physical process correlated with intensity are empirical matters. The experimental techniques are difficult (since they address phenomenal, rather than functional, consciousness), and they raise ethical issues (since they may require invasive procedures), but the questions have empirical content.

3.1.4 Protophenomenal Dependencies

What gives protophenomena their specific qualities? For example, why is one protophenomenen the experience of middle-C-here-now, but another the experience of pain-in-my-toe-now? Neuroscience suggests the answer, for there is no fundamental difference in the cortical tissues corresponding to middle-C-here-now and pain-in-my-toe-now. Rather, the qualitative nature of perceptions seems to depend on the connections between neurons. Several lines of evidence support this conclusion. First, there is the well-known phenomenon of referred pain (e.g., Karl, Birbaumer, Lutzenberger, Cohen & Flor, 2001), in which neurons reassign themselves from an amputated limb to another part of the body. Second, we may mention the very interesting experiments by Sur (2004), which demonstrate that neurons in auditory cortex can be made to support visual phenomena (i.e., perceptions that the animals experience as visual phenomena).

Neurophenomenological analysis allows us to transfer these observations from the neurophsychological domain to the theory of protophenomena. The implication is that protophenomena do not have inherent subjective qualities, but that they acquire them from interdependencies, which correspond to connections among neurons. (Thus protophenomenal theory may be classified as a *structural* theory of qualia.) As there are dynamical interdependencies among the activities of neurons, so there are necessarily correlated dependencies among the intensities of the corresponding protophenomena. From quantitative neurodynamical relationships we may hypothesize quantitative protophenomenal relationships (e.g., MacLennan, 1996b, App., 1999b), but we should not assume, in the absence of careful neurophenomenological investigation, that the subjective experience of protophenomenal intensity is proportional to a physical quantity at the activity site. In any case, neuroscience implies that the intensity of a single protophenomenon may depend on the intensities of tens of thousands (and sometimes hundreds of thousands) of other protophenomena and that their dynamical relationships can be highly complex (Anderson, 1995, p. 304).

Furthermore, the nervous system is a complex system, which exhibits a *macro-micro feedback loop* or *circular causality* (Solé & Goodwin, 2002, p. 150). This means that the behavior of the individual neurons, responding to their local environments, create a global state that in turn creates the neurons' local environments and thereby governs their behavior. Similarly, the intensities and dynamical relationships of the protophenomena create a global phenomenal state (a conscious state), which in its turn governs protophenomenal dynamics.

The ensemble of a person's protophenomenal intensities defines the degrees of freedom of their phenomenal world, and thus the universe of their possible conscious states. However, as we have seen, these intensities are not independent, but are constrained by dense and complex quantitative interrelationships, which therefore define the structure of that person's phenomenal world (their personal phenomenology). This structure is discovered through neurophenomenological experiment (i.e., mutually informing experimental phenomenology and experimental neuropsychology). Neurophenomenological analysis of protophenomenal dependencies provides an approach to explaining the specific qualities of phenomena (e.g., why sounds are experienced as sounds), the structure of phenomenal spaces, non-human perception, and spectral inversions, but they are not relevant to robot emotions and are discussed elsewhere (MacLennan, 1995, 1999a, 1999b, 2008a, 2008b).

3.1.5 Protophenomenal Change and Closure

Protophenomenal dependencies are not fixed, for they correspond to neural connections, which are altered by learning and other processes. Since protophenomenal dependencies define the structure of the phenomenal world, this too can change, with the result that phenomena may arise in consciousness that would not have before learning. Indeed, the phenomenal world changes its structure as the brain is restructured from infancy through adolescence in several waves of neuron proliferation and apoptosis (programmed death). There is evidence that even in adulthood neural proliferation takes place in some brain regions (e.g., Gould, Reeves, Graziano & Gross, 1999; Rakic, 2002), thus increasing the degrees of freedom of the phenomenal world.

A protophenomenal world is not causally closed, for there are protophenomena whose intensities do not depend on other protophenomena. The most obvious examples are sensory protophenomena, whose intensities depend on physical processes outside of the nervous system. Furthermore, processes outside the phenomenal world can alter protophenomenal dependencies; obvious examples are strokes, brain traumas, and degenerative diseases, which may permanently alter the structure of the phenomenal world. A third class of effects comes from alcohol and other psychoactive substances, which affect large numbers of protophenomenal dependencies, altering the dynamical relations among the protophenomena.

Since protophenomenal theory is not causally complete, but physical theory is generally supposed to be, the reader may wonder if protophenomenal theory is redundant. However, the body is an open system, and as the previous examples indicate, the nervous system is not causally complete, and this is reflected in the incompleteness of its phenomenal world. Further, quantum indeterminacy implies that, in an important sense, contemporary physics is causally incomplete. Most importantly, however, physical theory is fundamentally incomplete insofar as it does not explain subjective awareness, that is, insofar as the Hard Problem is not solved (denying the fact of consciousness is not a solution!). In particular, in the absence of something like protophenomenal theory, standard physical theory cannot answer the question of whether robots will feel their emotions (or, indeed, why we feel ours).

Protophenomenal analysis can be applied to many questions about consciousness, such as inverted qualia (spectral inversions), degrees of consciousness, consciousness in nonhuman animals, the unconscious mind, and the unity of consciousness, but these are not directly related to robot emotions, and so the reader is referred to other publications (MacLennan, 1995, 1996a, 1996b, 1999a, 1999b, 2008a, 2008b).

3.2 The Neurophenomenology of Emotion

A protophenomenological analysis of emotion involves qualitative and quantitative reductions of emotional experience, and correlates the results of these reductions with neurophysiological structures and processes. A qualitative reduction entails categorizing emotions (discussed below), but also identifying qualitatively different aspects of emotional experiences. For example, emotions have distinct aspects of appraisal, expression, and feeling (Plutchik, 2003, p. 287). One obvious quantitative aspect of an emotion is its intensity (we may be more or less happy, for example), but the quantitative aspects that are most relevant to protophenomenal reduction are those in which the emotion is extended in one or more spatial or abstract dimensions. For example, the feeling of fear includes the spatially extended, embodied sensations of perspiration, a throbbing heart, general trembling, tensed muscles, etc. (Panksepp, 2004, p. 207; Plutchik, 2003, p. 127).

A neurophenomenological analysis of emotion should begin with a phenomenology of emotion, that is, with an investigation of the structure of emotional experience, but the phenomenology of emotion is difficult (Plutchik, 2003, pp. 3–17, 64–7). People cannot always articulate the emotions they are feeling, they may misclassify their own emotions, and the classification of emotion differs between cultures and languages. People may intentionally misrepresent their emotional state (by verbal or nonverbal expression), misrepresent it to themselves, or repress it entirely (Plutchik, 2003, pp. 15–16). Indeed there are differences of opinion about whether a particular internal state even *is* an emotion (Plutchik, 2003, pp. 81–8). For example, Descartes classified wonder as an emotion, but Spinoza disagreed (Prinz, 2006, p. 87), and Plutchik (2003) classifies trust and anticipation as basic emotions (see below), but others question whether they are emotions at all (Prinz, 2006, p. 92). Notable forays into the phenomenology of emotion include Sartre (2001) and Hillman (1960).

While it is generally recognized that emotion can be subtle, nuanced, and elusive, it has been widely believed that there is a small number of basic emotions, of which the others are degrees, derivatives, cognitive elaborations, or mixtures (Prinz, 2004, pp. 86–94). However, at least since Descartes, philosophers and psychologists have proposed differing lists of basic emotions, often without much overlap (Hillman, 1960, p. 40). For example, Plutchik (2000, p. 64) lists anger / fear, anticipation / surprise, joy / sadness, and acceptance / disgust, while Ekman (1999) offers “amusement, anger, contempt, contentment, disgust, embarrassment, excitement, fear, guilt, pride in achievement, relief, sadness, satisfaction, sensory pleasure, and shame” (Prinz, 2004, p. 87). In part these difference are a consequence of differing criteria for determining basic emotions (Plutchik, 2003, pp. 69–72; Prinz, 2004, pp. 86–90).[3] For example, basic emotions can be classified on the basis of overt expression (e.g., Ekman, 1999), adaptive function (e.g., Plutchik, 2003, ch. 9), the neural systems that serve them (e.g., Panksepp, 2004), conceptual analysis (e.g., Gordon, 1987), or cross-cultural linguistic analysis (e.g., Wierzbicka, 1999). Prinz (2004, p. 90) advocates identifying basic emotions by “converging evidence” from several of these sources, which is a good basis for a neurophenomenological analysis.

Further, multidimensional scaling, latent semantic analysis, and similar statistical techniques have proved useful in exploring the phenomenological structure of emotions (Plutchik, 2003, chs. 4, 6). For example, Plutchik (2000) provides evidence that emotions have a three-dimensional structure and the topology of a cone. Based on subjects’ judgments of similarity and difference, words for emotions are mapped into two-dimensional space and lie in an approximate circle (called a *circumplex* of emotions; Plutchik & Conte, 1997). The circumplex represents two familiar aspects of emotional phenomenology: relative degrees of similarity and difference (e.g., distress is similar to misery), and bipolarity of affect (e.g., love vs. hate). On the basis of similarity, Plutchik divides the circumplex into eight segments corresponding to eight *basic* or *primary emotions*. Various *secondary emotions* are mixtures of the primary emotions. The analogy with primary and secondary colors on the color wheel is explicit, reflecting a frequent observation that colors and emotions have similar phenomenological structures (Plutchik, 2003, p. 103; cf. Gage, 1993; Goethe, 1840; MacLennan, 2008a, 2008b).

Another familiar aspect of emotional phenomenology is that emotions vary in intensity (e.g., annoyance < anger < rage). Intensity is represented by the third dimension of Plutchik’s emotion cone (its height): the maximum intensities of the emotions are around its circular base and intensity decreases towards its apex. All emotions have zero intensity and meet in emotional neutrality at the apex. Plutchik (1980) notes that the three emotional aspects of similarity, polarity, and intensity correspond to properties of color: hue, complementarity, and intensity, and therefore that the emotional cone is analogous to the color sphere (more, precisely, to its lower hemisphere). Plutchik’s map of the emotions has been criticized (e.g., Prinz, 2004, pp. 92–3), because some of his primary emotions do not seem to be basic, and because some higher cognitive emotions (e.g., romantic jealousy) seem to be more than blends of primary emotions, but it is an important investigation in the phenomenology of emotion.

Valence, that is, whether an emotion is positive or negative (desirable or undesirable, and thus leads to positive or negative reinforcement, and approach or avoidance), is commonly assumed to be a fundamental dimension in the phenomenology of emotions (Panksepp, 2004, p. 46; Plutchik, 3002, p. 21; Prinz, 2004, pp. 167–78; Rolls, 2007). However, even this simple distinction is problematic; for example, nowadays both anger and pride are sometimes viewed as positive emotions, although traditionally they were considered negative (Plutchik, 2003, pp. 7–8). Further, anger, although considered negative in being un-

[3] Indeed, the concept of a basic emotion has been criticized (e.g., Griffiths, 1997; Ortony & Turner, 1990).

pleasant, is positive in terms of approach (i.e., attack); fear may involve both an avoidance of the fearful stimulus, but also an active approach toward safety (Prinz, 2004, p. 168). Movie-goers seek to feel ostensibly negative emotions such as fear and anger (DeLancey, 2002, pp. 7–8).

Fortunately there is a large body of research on the neurophysiology of emotions in the context of evolutionary biology (e.g., Panksepp 2004; Plutchik 2003; Prinz 2004). An evolutionary approach has the advantages of not limiting attention to human emotion and of considering the adaptive functions of emotion, both of which are important to the issue of robot emotion. By understanding how emotions fulfill essential functions in nonhuman animals, we can begin to understand how to design synthetic emotions in robots to fulfill analogous or even completely different functions. Such emotions may serve goals such as performing physical tasks (e.g., mining, construction, agriculture), collecting information or samples, cleaning up hazardous materials, rescuing or protecting humans or other animals, responding to accidents and medical emergencies, and achieving military objectives.

Another potentially important aspect of emotional phenomenology is the recent discovery of *mirror neurons*, which mimic the activity of neurons in another animal based on unconscious perception of that other animal (Arbib & Rizzolatti, 1997; Rizzolatti & Craighero, 2004; Rizzolatti, Fadiga, Gallese & Fogassi, 1996). Mirror neurons thus provide a mechanism by which the emotional state of one animal may be subconsciously internalized into the nervous system of the other, providing a neural basis for empathy. Mirror neurons seem to have a number of functions that may also be valuable in robots, including the facilitation of cooperative behavior and learning by imitation (Breazeal et al., 2004), as well as "reading the mind" of another agent by mirroring its neural state.

3.3 Possible Biological Correlates of Emotional Protophenomena

Next I will discuss the conditions under which humans *feel* their emotions. It may seem axiomatic that we feel our emotions, but that is not so; in fact it is important to distinguish emotion and feeling (Damasio, 1999, pp. 42–9). The primary emotional response takes place in the limbic system and is unconscious, whereas conscious emotional experience is confined to cortical areas. Within the limbic system, the amygdala can respond to sensory information routed through the thalamus to initiate an appropriate emotional response to the stimulus. This response may be innate or based on memories stored in the amygdala (one basis for unconscious conditioned responses). The amygdala also generates emotional responses on the basis of cortical processing of information from the senses or memory when that processing determines that an emotional response may be required. In either case the amygdala then sends out signals (via the hypothalamus) to initiate an appropriate response (e.g., fight or flight, care-giving), which includes neuromotor readiness and initiation, activation of the sympathetic nervous system, hormonal secretions (e.g., adrenaline), and changes in heart rate, breathing, perspiration, etc. This process is unconscious.

These changes in the physiological state of the body are relayed (back through the hypothalamus) to a variety of cortical areas, where they are consciously experienced as emotion. The felt emotion is a complex product of these physiological factors, of the consciously perceived and interpreted stimulus or memory, and of other contextual factors.

Along with sensory information from multimodal sensory-integration areas, such as the tectum, the visceral and physiological information is integrated in convergence areas (such as the periaqueductal gray region of the midbrain), which create the conscious experience of a unified organism and its emotional-behavioral state (Damasio, 1999, pp. 60–1). The resulting conscious cognition of the emotional state can modulate the activity of the amygdala, reinforcing it if the emotional response is perceived to be appropriate, or dampening it down if not.

William James (1884) and Carl Lange (1885) independently made the surprising claim that the foundation of emotional experience is sensation of prior bodily change (the *James-Lange theory* or *somatic feeling theory*). Although there have been objections, an increasing body of neuropsychological data supports various modifications and extensions of his theory (e.g., Damasio, 1994, 1999; Prinz, 2004). Prinz (2004, ch. 9) suggests that a three-level *emotional processing hierarchy* underlies emotional consciousness, a view that is quite compatible with protophenomenal analysis. At the lowest level are neurons (in primary somatosensory cortex, pons, insula) with small receptive fields responding to local conditions in skeletal muscles, visceral organs, hormone levels, etc.; these correspond to emotional protophenomena. At the intermediate level neurons (in secondary somatosensory, dorsal anterior cingulate, and insular cortices) integrate the protophenomena into coherent patterns of activity, that is, into emotional phenomena. These seem to be similar to the *first-* and *second-order maps* described by Damasio (1999, ch. 6). At the third level these patterns are characterized and specific emotions are recognized, perhaps in ventromedial prefrontal cortex and rostral anterior cingulate cortex (Prinz, 2004, p. 214). Prinz's hierarchy can also be compared to Damasio's (1999, ch. 9) three-level hierarchy of emotion, feeling, and feeling feeling: "*an emotion*, *the feeling of that emotion*, and *knowing that we have a feeling of that emotion*" (Damasio, 1999, p. 8, italics in original). Thus, like other conscious phenomena, the qualitative character of an emotional phenomenon consists in the interdependencies among its constituent protophenomena.

Because emotions can trigger the release of neuromodulators that affect the activity of large groups of neurons (e.g., Damasio, 1999, pp. 281–2; Fellous, 1999), there is a parallel effect on large ensembles of protophenomenal interdependencies. This alters the dynamics of the protophenomena, which is experienced as a change in the conscious process. Emotionally-triggered neuromodulation affects neurodynamics so that it better serves the function of the emotion, and it is reasonable to suppose that autonomous robots will also benefit from emotion-triggered pervasive alterations to their cognitive processes. Further, the relatively global and non-specific effects of these neuromodulators permit the physical characteristics of activity sites to be investigated with less invasive procedures than would otherwise be required.

Therefore, to obtain a comprehensive explanation of the structure of conscious emotional experience, we need to investigate the representation and integration of information in these cortical areas, especially those aspects related to emotional response, and to correlate these neuropsychological investigations with phenomenological investigations into the structure of conscious emotional experience. Such investigations are in progress. For example, Damasio, Grabowski, Bechara, Damasio, Ponto, Parvizi, and Hichwa (2000) have identified some of the different cortical and subcortical brain regions involved in various consciously experienced emotions. As a consequence of studies such as these we will be able to identify the neuronal processes correlated with emotional protophenomena, and the interdependencies among them that define the qualitative structure of felt emotion. This is, of course, an ongoing and long-term research project, but neurophenomenological research into human emotional experience already provides a basis for understanding the determinants of the phenomenology of robot emotion.

3.4 The Protophenomena of Robot Emotion

Based on the forgoing analysis of conscious and unconscious emotional response in humans, we can address the problem of conscious emotional response in robots in a more focused way. Protophenomena are elementary subjective degrees of freedom, which correspond to activity sites in the brain, so that physical processes at these sites are correlated with the presence of the corresponding protophenomena in conscious experience. Is it possible that physical processes in a robot's "brain" (central information processor) could constitute activity sites with associated emotional protophenomena? The issue is whether the robot's information processing devices are sufficiently similar to the human brain's *in the relevant ways*. For example, if the robot's processor were an actual living human brain, and the robot's body were sufficiently similar to a human's, there would seem to be no reason to deny that it was having genuine subjective experiences. On the other hand, there may be something about living neurons that makes them the only physical systems capable of supporting consciousness. An analogy may clarify the issue. Water is liquid, but its liquidity depends on physical properties of H_2O molecules (such as their finite volume and short-range mutual attraction), which are also possessed by some other molecules. Therefore, there are other liquids besides water or, to put it differently, liquidity can be realized by a variety of substances. So the questions are: What are the properties by virtue of which "neural stuff" supports protophenomena? Are there other non-neural, physical systems that have these properties, or can we make them?

Unfortunately, at this stage of the scientific investigation of consciousness we cannot say what properties of physical systems are sufficient for them to be activity sites and support protophenomena. Nevertheless, the question is empirical, since it can be addressed by controlling physical quantities and substances in individual neurons and observing their effects on conscious experience. The technology for conducting these experiments is improving. For example, Losonczy, Makara, and Magee (2008) have developed techniques for delivering individual neurotransmitter molecules to individual dendritic spines with a spatial resolution of 1 micrometer and time resolution of 1 millisecond (but they were not applied *in vivo*). Therefore we anticipate that it is just a matter of time before we have a better understanding of the essential physical properties of activity sites. In the meantime we may consider several plausible possibilities.

First, the activity sites could be the somatic membranes of neurons and protophenomenal intensity might correspond to membrane potential relative to its resting potential. Before we could decide whether similar activity sites could be constructed in an artificial system, we would need to have a more detailed understanding of the relation of membrane potential to protophenomenal intensity. For example, must it be the membrane potential of a living cell, or could a nonliving membrane support a protophenomenon? Is it purely an electrical property, or an electrochemical one? Does it depend on specific ions or on a particular membrane structure (e.g., lipid bilayer)? These are all empirical questions, and their answers will delimit the sorts of artificial physical devices that could support protophenomena.

If, as Cook (2000, 2002a, 2002b, chs. 6–7, 2008) suggests, the intensity of a protophenomenon correlates with the flux of ions across the cell membrane when the ion channels open during an action potential, and the protophenomenon is in effect the cell's sensing of its (intercellular) environment, then the essential properties of an activity site might include a boundary separating it from its environment, the ability to sense its environment, and the ability to modify the environment as a consequence. In this case, it would seem to be possible to construct an artificial device supporting protophenomena, but the specific requirements would have to be determined empirically.

Chalmers (1996, ch. 8) considers the possibility that *information spaces* may provide the link between the physical and the phenomenal, since they can be realized either as physical systems or as phenomenological structures. In particular, he suggests that quite simple physical systems might have associated protophenomena (p. 298). An information space is characterized by "differences that make a difference," that is, by distinctions that causally affect behavior. The physically realized information space must have a sufficient number of states to support the distinctions and must have the appropriate causal relations. (That is, like general information processing systems, there is a homomorphism from the physical realization to the information system, i.e., the physical system has at least the abstract structure of the information system, but may have additional structure irrelevant to the information processing: MacLennan, 1994, 2004.) The structure of the phenomenal space corresponds to the structure of the information space (due, in protophenomenal terms, to the protophenomena having interdependencies that correspond to the causal relations in the physical system).

Further, Chalmer's hypothesis and Cook's theory seem to be compatible. The binding of neurotransmitters to their receptors conveys information to a neuron about its extracellular environment, which can be quantified as an increase in the *system mutual information* between the cell and its environment. Each receptor make a contribution measured by the *conditional mutual information* of the resulting postsynaptic state.[4] This is a process of input transduction, which converts the many different neurotransmitters and ions in the region of the synapse into a common computational currency, membrane potential. These electrical signals propagate down the dendritic tree to combine in the somatic membrane potential, which thus accomplishes a simple form of sensor integration. Much of this process is electrically passive and nearly linear (neglecting some voltage gated channels). Under appropriate circumstances, such as the membrane potential at the axon hillock exceeding a threshold, an action potential is generated, which is a highly nonlinear and active process; it has the character of an elementary decision (e.g., the level of excitation is above the threshold). The resulting action potential causes specific chemicals (neurotransmitters) to be released at the axon terminal, which is a kind of output transduction, and constitutes the action resulting from the decision (the difference that makes a difference). Thus, the neuron can be viewed as a simple control system, with input sensors, some simple information processing resulting in a decision, and output effectors or actuators. This corresponds to Cook's (2000, 2002a, 2002b, chs. 6–7, 2008) idea that cognition

[4] See, for example, Hamming (1980, §7.6) for definitions of these terms.

(information processing) is associated with the synaptic processes and consciousness (protophenomena) with the action potential. A spike in protophenomenal intensity would occur with the generation of an action potential. Such a scenario suggests that protophenomena might be associated with other simple control systems, in which inputs are translated into a computational medium and integrated to trigger a decision (an active nonlinear process) in order to have some physical effect. It doesn't seem to be impossible that nonbiological control systems of this kind would have associated protophenomena, but of course it is an empirical question.

In summary, if Chalmer's suggestion is correct, then many physically realized information spaces will be activity sites with associated protophenomena. In particular, since a robot's processor is devoted to the physical realization of information spaces, it would be reasonable to suppose that its constituent devices would have associated protophenomena. This would not, of course, imply that the robot is conscious, for protophenomena are not yet phenomena, but if the information processing were organized to create the appropriate protophenomenal interdependencies so that they cohered into phenomena and created a phenomenal world, then we could say that the robot is conscious. In particular, an appropriate structure among the protophenomena would produce emotional phenomena (felt emotions).

3.5 The Combinatorial Structure of Robot Emotion

We have seen that it is an empirical matter what sorts of physical objects have associated protophenomena, and therefore whether the components of a robot could support protophenomena. However, protophenomena are necessary, but not sufficient, for consciousness, which also requires that the protophenomena have relations of interdependency sufficient for the emergence of phenomena. Therefore, in the case of robot feelings we must consider the structure of artificial emotional phenomena.

In robots, as in animals, a primary function of emotion is to make rapid assessments of external or internal situations and to ready the robot to respond to them with action or information processing. This may involve power management, shifting energy to more critical systems, adjustment of clock rates, deployment and priming of specialized actuators and sensors, initiation of action, and so forth. These processes will be monitored by *interoceptors* (internal sensors) that measure these and other physical properties (positions, angles, forces, stresses, flow rates, energy levels, power drains, temperatures, physical damage, etc.) and send signals to higher cognitive processes for supervision and control. Therefore, many of these interoceptors will be distributed around the robot's body and this spatial organization will be reflected in somatosensory maps or other information structures. As a consequence patterns among the interoceptive signals will be represented, and the associated protophenomena will cohere into spatially organized phenomena.

In this way emotional phenomena are structured spatially in relation to the body, but these phenomena also have a qualitative structure, which may vary depending on the input space of the interoceptors. Each interoceptor will have a response curve defined over its input space, but connections among the interoceptors at a location and connections to higher-order sensory areas will stitch together a topology representing the joint input space (MacLennan, 1995, 1999b). This topology defines the qualitative structure of the resulting emotional phenomena.

Some of a robot's sensory spaces will be similarly structured spatially and qualitatively to ours, and in these cases we can expect the robot's emotional experiences (its *feelings*) to be similar to our own. Examples might include pressure sensors in the skin and angle and stress sensors in the joints. On the other hand, other interoceptors will be quite different from humans'. For example, a robot is unlikely to experience a quickened heartbeat or shallow, rapid breathing (because it is unlikely to have a heart or lungs), and we, in contrast, do not experience a redistribution of electrical power, which a robot might. These interoceptive spaces will have their own topologies, which determine their phenomenal structures, and so in these cases we must expect the robot's emotional experiences to be significantly different and alien to us. Although we may be unable to imagine them, we will be able to understand their abstract structure, which will give us some insight into the robot's experience. In general, a robot's emotions will be peculiar to its "form of life," as ours are to ours. (See MacLennan (1996a) for more on understanding non-human perception.)

If this is the case, one might question why these robotic experiences should be considered emotions at all. One reason is their similar function to natural emotions (recall **Section 2 Background**). For example, they will reflect general goals of critical importance to the robot's behavior, which are therefore directly motivating, and that consequently have a persisting, pervasive, appropriate effects on the physical state of the robot (by controlling sensors, effectors, and information processing). Another reason is that, due to the need for rapid, pervasive response, these experiences will have unconscious roots (i.e., below the level of coherent phenomena); conscious experience will be secondary and modulated by the already activated emotion.

4 Future Trends

As explained above, we cannot determine whether it would be possible for a robot to have feelings (emotional experiences), but there is every reason to believe we will be able to decide in the future. The answer depends on the results of two investigations. First, we will need to understand what physical processes support protophenomena, that is, what sorts of physical systems can be activity sites. This will require a detailed investigation of neurons in order to determine what physical (or biological) characteristics are necessary and sufficient for protophenomena. Such experiments will involve *in vivo* manipulation of structures and substances in individual neurons associated with *salient protophenomena* (protophenomena that individually or in small groups constitute phenomena, i.e., are experienceable). Although the procedures will be difficult (and will raise ethical issues), there is rapid progress and so it seems likely that these questions will eventually be answered empirically.

The results of these experiments will allow us to make plausible hypotheses about the necessary and sufficient conditions for protophenomena, and therefore about whether any particular artificial system could support them. Further experiments could test these hypotheses (e.g., by replacing, in a living animal, neural structures or processes by artificial surrogates). As a result we will be able to make experimentally verified statements about the sorts of physical systems that support protophenomena, and therefore the sorts of robot technology that could support consciousness.

However, as we have seen, protophenomena are not sufficient in themselves for feeling emotions; it is necessary that the protophenomena be so structured, through their interdependencies, to cohere into emotional phenomena. An improved neurophenomenological understanding of human and animal emotional experience, especially as implemented in the cortex, will show us how to interconnect robotic activity sites so that emotional phenomena emerge. The experiments in this case are not so difficult, but there is much we still do not understand about the cortical experience of emotion, but this will come in time with progress in neuroscience and neurophenomenology.

I hope that I have convinced you that protophenomenological analysis allows the question of robot feelings to be investigated empirically. Yet the breadth and depth of the investigations is daunting, and you may wonder whether they are worth the effort. Why should we care whether robots might be able to feel their emotions?

First, and perhaps paradoxically, these investigations will help us understand ourselves and other animals. Unless we can give a principled answer to the question of robot feelings (whether positive or negative) there will be a serious gap in our understanding of our own humanity. That is, so long as we cannot say why robots could or could not feel their emotions, we cannot really explain why we feel ours. Since emotions are essential to human nature, this issue is fundamental to our self-knowledge.

Another, more distant reason for investigating robot feelings is the matter of robot rights. If we are eventually able to build robots with intelligence comparable to humans, then they will have an emotional system of comparable complexity (although, perhaps, very different in its particulars), because indeed emotion is essential to intelligence. Rights (animal as well as human) often presuppose the capacity to suffer, and so robot rights (in general and in specific cases) might depend on whether they can feel their emotions.

5 Conclusions

In conclusion, I have argued that it is by no means impossible that some future robots may feel their emotions, that is, that they may have subjective emotional experiences homologous, but not identical, to ours. To determine the precise conditions sufficient for robot feelings it will be necessary to conduct detailed neurophenomenological investigations of subjective experience in order to isolate the physical processes correlated with the smallest units of that experience. This will enable us to formulate empirically testable hypotheses about the sorts of nonliving physical systems (if any) that may support protophenomena, and therefore conscious experience. This, in itself, is not sufficient to imply that robots could feel their emotions, for it is also necessary to understand neural structures underlying emotional experience, and the corresponding interdependencies among emotional protophenomena. (The emotional protophenomena are not, of course, independent of other protophenomena, such as those associated with bodily sensation.) The results of these neurophenomenological investigations will show us how to structure the emotional protophenomena of robots so that they cohere into emotional phenomena, that is, so that the robots feel their emotions. Thus, although significant unanswered questions remain, they can be addressed empirically, and their answers will allow us to decide whether robots could feel their emotions.

6 References

Anderson, J.A. (1995). *An introduction to neural networks*. Cambridge, MA: MIT Press.

Arbib, M.A., & Rizzolatti. R. (1997). Neural expectations: A possible evolutionary path from manual skills to language. *Communication and Cognition*, 29, 393–423.

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

Breazeal, C. (2003). Toward sociable robots. *Robotics and Autonomous Systems*, 42, 167–75.

Breazeal, C., Brooks, A., Gray, J., Hoffman, G., Kidd, C., Lee, H., Lieberman, J., Lockerd, A., & Chilongo, D. (2004). Tutelage and collaboration for humanoid robots. *International Journal of Humanoid Robots*, 1(2), 315–48.

Campbell, K.K. (1970). *Body and mind*. New York, NY: Doubleday.

Chalmers, D.J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2, 200–19.

Chalmers, D.J. (1996). *The conscious mind*. New York, NY: Oxford University Press.

Chalmers, D.J. (2002). Consciousness and its place in nature. In D.J. Chalmers (Ed.), *Philosophy of mind: Classical and contemporary readings*. Oxford, UK: Oxford.

Cook, N.D. (2000). On defining awareness and consciousness: The importance of the neuronal membrane. In *Proceedings 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 T. Crow (Ed.), *The speciation of modern Homo sapiens*. London, UK: Proceedings of the British Academy, v. 106 (ch. 9).

Cook, N.D. (2002b). *Tone of voice and mind: The connections between intonation, emotion, cognition and consciousness*. Amsterdam, Netherlands: John Benjamins.

Cook, N.D. (2008). The neuron-level phenomena underlying cognition and consciousness: Synaptic activity and the action potential. *Neuroscience*, 153(3), 556–70.

Damasio, A.R. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York, NY: Avon.

Damasio, A.R. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. New York, NY: Harcourt, Brace & Co.

Damasio, A.R., Grabowski, T.J., Bechara, A., Damasio, H., Ponto, L.L.B., Parvizi, J., & Hichwa, R.D. (2000). Subcortical and cortical brain activity during the feeling of self-generated emotions. *Nature Neuroscience*, 3, 1049–56.

DeLancey, C. (2002). *Passionate engines: What emotions reveal about mind and artificial intelligence*. Oxford, UK: Oxford University Press.

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

Ekman, P. (1999). Basic emotions. In T. Dalgleish & T. Power (Eds.), *The handbook of cognition and emotion* (pp. 45–60). New York, NY: Wiley.

Fellous, J.-M. (1999). Neuromodulatory basis of emotion. *The Neuroscientist*, 5(5), 283-94.

Gage, J. (1993). *Color and culture: Practice and meaning from antiquity to abstraction*. Boston, Toronto, & London: Little, Brown, & Co.

Goethe, J. W. von. (1840). *Goethe's theory of colours* (C.L. Eastlake, tr.). London, UK: Murray.

Gordon, R. (1987). *The structure of emotions*. Cambridge, UK: Cambridge University Press.

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

Gregory, R.L. (ed.) (1987). The Oxford companion to the mind. Oxford, UK: Oxford University Press.

Hamming, R.W. (1980). *Coding and information theory*. Englewood Cliffs, NJ: Prentice-Hall.

Hillman, J. (1960). *Emotion: A comprehensive phenomenology of theories and their meanings for therapy*. Evanston, IL: Northwestern Univ. Press.

Ihde, D. (1986). *Experimental phenomenology. An introduction*. Albany, NY: State University of New York Press.

James, W. (1884). What is an emotion? *Mind*, 9, 188–205.

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

Kirk, R. (1974). Zombies versus materialists. *Aristotelian Society*, 48 (suppl.), 135–52.

Kleinginna, P.R., & Kleinginna, A.M. (1981). A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation & Emotion*, 5(4), 345–79.

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

Kripke, S.A. (1980). *Naming and necessity*. Cambridge, MA: Harvard University Press.

Lange, C.J. (1885). *Om sindsbevaegelser: Et psyko-fysiologisk studie*. Copenhagen: Jacob Lunds.

Laughlin, C.D., Jr., McManus, J., & d'Aquili, E.G. (1990). Brain, symbol and experience: Toward a neurophenomenology of consciousness. Boston, MA: New Science Library.

Llinas, R.R. (1988). The intrinsic electrophysiological properties of mammalian neurons. *Science*, 242, 1654-64.

Losonczy, A., Makara, J.K., & Magee, J.C. (2008). Compartmentalized dendritic plasticity and input feature storage in neurons. *Nature*, 452, 436–40.

Lutz, A., & Thompson, E. (2003). Neurophenomenology: Integrating subjective experience and brain dynamics in the neuroscience of consciousness. *Journal of Consciousness Studies*, 10(9/10), 31–52.

Lyons, W. (1986). *The disappearance of introspectionism*. Cambridge, MA: MIT Press.

MacLennan, B.J. (1994). Continuous computation and the emergence of the discrete. In K.H. Pribram (Ed.), *Rethinking neural nets: Quantum fields and biological data* (pp. 199–232). Hillsdale, NJ: Lawrence-Erlbaum.

MacLennan, B.J. (1995). The investigation of consciousness through phenomenology and neuroscience. In J. King & K.H. Pribram (Eds.), *Scale in conscious experience: Is the brain too important to be left to specialists to study?* (pp. 25–43). Hillsdale, NJ: Lawrence Erlbaum.

MacLennan, B.J. (1996a). The elements of consciousness and their neurodynamical correlates. *Journal of Consciousness Studies*, 3 (5/6), 409−24. Reprinted in J. Shear

(Ed.), *Explaining consciousness: The hard problem* (pp. 249–66). Cambridge, MA: MIT, 1997.

MacLennan, B.J. (1996b). *Protophenomena and their neurodynamical correlates* (Technical Report UT-CS-96-331). Knoxville, TN: University of Tennessee, Knoxville, Department of Computer Science. Available: www.cs.utk.edu/~mclennan

MacLennan, B.J. (1999a). Neurophenomenological constraints and pushing back the subjectivity barrier. *Behavioral and Brain Sciences*, 22, 961–63.

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

MacLennan, B.J. (2003). Color as a material, not an optical, property. *Behavioral and Brain Sciences*, 26, 37–8.

MacLennan, B.J. (2004). Natural computation and non-Turing models of computation. *Theoretical Computer Science*, 317, 115–45.

MacLennan, B.J. (2008a). Consciousness: Natural and artificial. *Synthesis Philosophica*, 22(2), 401–33.

MacLennan, B.J. (2008b). Protophenomena: The elements of consciousness and their relation to the brain. In A. Batthyány, A. Elitzur & D. Constant (Eds.), *Irreducibly conscious: Selected papers on consciousness* (pp. 189–214). Heidelberg & New York: Universitäts-verlag Winter.

McCall, R.J. (1983). *Phenomenological psychology: An introduction. With a glossary of some key Heideggerian terms*. Madison, WI: University of Wisconsin Press.

Ortony, A., & Turner, W. (1990). What's basic about basic emotions? *Psychological Review*, 97, 315–31.

Panksepp, J. (2004). *Affective neuroscience: The foundations of human and animal emotions*. New York, NY: Oxford University Press.

Plutchik, R. (1980). *Emotion: A psychoevolutionary synthesis*. New York, NY: Harper & Row.

Plutchik, R. (2000). *Emotions in the practice of psychotherapy: Clinical implications of affect theories*. New York, NY: American Psychological Association.

Plutchik, R. (2003). *Emotions and life: Perspectives from psychology, biology, and evolution*. New York, NY: American Psychological Association.

Plutchik, R., & Conte, H.R. (Eds.). (1997). *Circumplex models of personality and emotions*. Washington, DC: American Psychological Association.

Prinz, J. (2006). *Gut reactions: A perceptual theory of emotion*. New York, NY: Oxford University Press.

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

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

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3(2), 131–41.

Rolls, E.T. (2002). Emotion, neural basis of. In N.J. Smelsner & P.B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (pp. 4444–9). Amsterdam, Netherlands: Pergamon.

Rolls, E.T. (2005). *Emotion explained*. Oxford, UK: Oxford Univ. Press.

Rolls, E.T. (2006). Brain mechanisms of emotion and decision-making. *International Congress Series*, 1291, 3–13. Amsterdam: Elsevier.

Rolls, E.T. (2007). A neurobiological approach to emotional intelligence. In G. Matthews, M. Zeidner & R.D. Roberts (Eds.), *The science of emotional intelligence* (pp. 72–100). Oxford, UK: Oxford Univ. Press.

Rudrauf, D., Lutz, A., Cosmelli, D., Lachaux, J.-L., & Le Van Quyen, M. (2003). From autopoiesis to neurophenomenology: Francisco Varela's exploration of the biophysics of being. *Biological Research*, 36(1), 27–65.

Sartre, J. (2001). *Sketch for a theory of the emotions*. New York, NY: Routledge.

Solé, R., & Goodwin, B. (2002). *Signs of life: How complexity pervades biology*. New York, NY: HarperCollins Publishers.

Sur, M. (2004). Rewiring cortex: Cross-modal plasticity and its implications for cortical development and function. In G.A. Calvert, C. Spence & B.E. Stein (Eds.), *Handbook of multisensory processing* (pp. 681–94). Cambridge, MA: MIT Press.

Varela, F.J. (1996). Neurophenomenology: A methodological remedy to the hard problem. *Journal of Consciousness Studies*, 3, 330–50. Reprinted in: J. Shear (Ed.), *Explaining consciousness: The hard problem of consciousness* (pp. 337–58). Cambridge, MA: MIT Press, 1997.

Wierzbicka, A. (1999). *Emotions across languages and cultures: Diversity and universals*. Cambridge, UK: Cambridge University Press.

7 Key Terms and Definitions

7.1 Action Potential

An action potential, also called a *neural impulse* or *spike*, is a stereotypical excursion in the membrane potential caused by the opening and closing of ion channels in a cycle of positive feedback and recovery. An action potential is triggered when a membrane is sufficiently depolarized from its normal negative resting potential; positive feedback causes a rapid repolarization in the positive direction, after which there is a relatively slow return to a potential slightly more negative than the resting potential, followed by a gradual return to resting potential. Action potentials propagate down axons without attenuation to convey information to other neurons.

7.2 Axon Hillock

The axon hillock is the base of an axon, that is, the region of a neuron's soma (q.v.) from which the axon projects. In many neurons it is the place where action potentials (q.v.) are generated.

7.3 First-person

In the context of consciousness studies, *first-person* refers to the experience of one's own consciousness. In contrast to *third-person* observation (q.v.), the observer is not separable from the observed. Such observation is inherently *private*, but the techniques of neurophenomenology (q.v.) permit the establishment of an observer-independent body of *public* fact on which scientific theories can be built.

7.4 Hard Problem

The "Hard Problem" is the term introduced by Chalmers (1995) to refer to the principal problem of the scientific investigation of consciousness, namely, the integration of the primary fact of conscious experience with contemporary scientific understanding of the material universe.

7.5 Neurophenomenology

Neurophenomenology combines the phenomenological (q.v.) investigation of the structure of experience with the neuroscientific investigation of the neural correlates of that experience. Thus is promises a coherent account of experience from both first-person (q.v.) and third-person (q.v.) perspectives.

7.6 Objective

Objective and *subjective* (q.v.) are used to make two different distinctions, which overlap, but confusion between the distinctions muddies the mind-body problem. In the context of this chapter, *objective* refers to a *third-person* (q.v.) perspective, as opposed to a *subjective* or *first-person* perspective (q.vv.). Colloquially, *objective* connotes the unbiased, factual, and scientific, but that is not the meaning here, since phenomenology (q.v.) seeks unbiased, factual, and scientific knowledge based on subjective observation.

7.7 Phenomenology (Phenomenological)

Phenomenology, especially as developed by Husserl, Heidegger, and Merleau-Ponty, is the systematic investigation of the invariant structure of experience by empirical, but first-person (q.v.) methods. Accurate phenomenology requires systematic training, which distinguishes from naive introspection.

7.8 Protophenomenological Analysis

Protophenomenological analysis seeks to explain the structure of conscious experience in terms of the interdependencies among protophenomena (q.v.) as determined by neurphenomenology (q.v.).

7.9 Protophenomenon

Protophenomena are the smallest units of conscious experience, which are hypothesized and investigated on the basis of neurophenomenological research (q.v.), that is, on the basis of coordinated phenomenology (q.v.) and neuroscience.

7.10 Qualia

Qualia (singular: *quale*) are the felt qualities of phenomena, as aspects of first-person (q.v.) or subjective (q.v.) experience. Examples of qualia are the feeling of warmth of a warm thing, the auditory experience of a C-major chord, the feeling in the gut of anger or fear, and so forth.

7.11 Soma

The soma is the cell body of a neuron. Inputs to a neuron causes fluctuations in the electrical potential across the neuron membrane, which are integrated into the somatic membrane potential. In a typical neuron, and to a first approximation, a sufficiently large depolarization of the membrane at the axon hillock (q.v.) will trigger the generation of an action potential (q.v.).

7.12 Subjective

There are two distinct but overlapping senses in which something may be termed *subjective* and contrasted with the *objective* (q.v.). In the context of this chapter, *subjective* refers to first-person (q.v.) observation, which is essential to the protophenomenological analysis (q.v.) of emotion. Colloquially, *subjective* may connote observations and opinions that are biased or distorted, but that is not the intent here, since the purpose of phenomenology (q.v.) is to produce unbiased and factual first-person (subjective) observations.

7.13 Third-person

In the context of consciousness studies, *third-person* is used to refer to ordinary scientific observation of some object separate from the observer. For example, we may make third-person observations of some physical system, of the brain, or of some person's behavior (including verbal report). Third-person observation can be a public process grounded in shared observational practices leading to a provisional consensus about observed facts.

Often taken to be synonymous with *objective* (q.v.) and contrasted with *first-person* and *subjective* (q.vv.).