Biophysics of Consciousness: A Foundational Approach
R. R. Poznanski, J. A. Tuszynski and T. E. Feinberg

3

HOW LANGUAGE EVOLUTION RESHAPED HUMAN CONSCIOUSNESS

Michael A. Arbib

The ABLE Project: Action, Brain, Language, Evolution
University of Southern California
Los Angeles CA 90089-2520, USA
arbib@usc.edu

Abstract

This chapter discusses the view that human consciousness may share aspects of "animal awareness" with other species, but has its unique form because humans possess language. Four ingredients of a theory on the evolution of human consciousness are offered: the view that a précis of intended activity is necessarily formed in the brain of a human that communicates in a human way; the notion that this defines a "communication plexus" essential to a particularly human form of consciousness; the mirror system hypothesis for the evolution of the human language-ready brain, and a view of the co-evolution of subtleties of emotion atop more basic systems for motivation such as thirst, hunger, fear and mating.

Keywords: Consciousness; animal awareness; language; evolution; brain; communication plexus; mirror system hypothesis; language-ready brain; emotion; motivation.

3.1. Introduction

The "core theory" here is what I will call the A & H theory — the account by Mary Hesse and myself (Arbib, 1985; Arbib & Hesse, 1986) of how human consciousness evolved in a way that gave a central but not exclusive role to language in our conscious experience. (When pain dominates our consciousness, the word "pain" is inessential.) A subsequent paper (Arbib, 2001) offered a first attempt to integrate this theory with the mirror system hypothesis of the evolution of the language-ready brain (Arbib & Rizzolatti, 1997; Rizzolatti & Arbib, 1998), a hypothesis which has been modified and expanded across the years. This paper reproduces, edits and re-organizes portions of the 2001 account, while also paying more attention to the role of emotion in consciousness, assessing the relevance of comparative accounts of monkeys, apes and humans (comparative neuroprimatology), and briefly relating our approach to other brain-related accounts of consciousness.

Section 2 contrasts brain states from mental states, stressing that the mental state is a gross abstraction of the brain state and is at *approximately* the same level as language in which a mental state may *or may not* be expressible in words, while the level of detail that we ascribe to a mental state is far coarser than a brain state. This serves as the basis for expressing our view of what an account of consciousness has to explain.

3.2. Neurology and Consciousness

One of the definitions of *consciousness* in the Oxford English Dictionary is: "The … capacity from which awareness of thought, feeling, and volition and of the external world arises; the exercise of this. … [Also] the aspect of the mind made up of operations which are known to the subject." We share, I suggest, some aspects at least of awareness of feelings concerning the external world with other species (perhaps especially with mammals) — I call this "animal awareness" — but humans are unique in that such awareness is so often accompanied by awareness of thought, volition and at least a folk theory of the mind. Of particular interest, here, is the way our conscious experience as humans is accompanied by a "running commentary" which makes language so much a part of our consciousness and of our ability to share at least some aspects of it with others. The etymological basis is the Latin *con* + *scio* — "to know *with*" and I suggest that it is the ability to share our experience *with others* that is the defining feature of consciousness, even though it remains active when we are alone. Dreaming would seem to be a further etiolation of consciousness — as some processes that support our conscious interaction with the physical and social world may continue even when our sensory and motor processes are in abeyance.

The prime aim of this paper is to place a theory of consciousness from the 1980s (Arbib, 1985; Arbib & Hesse, 1986) in the perspective offered by my recent thinking about language and human emotion in an evolutionary perspective strongly influenced by neuroscience. But first, I offer a few examples from the neurological literature to suggest some subtleties in linking consciousness to the brain. First is the phenomenon of *blindsight* which demonstrates that neural activity can effectively link perception and action without intervention of consciousness. Neurologists long held that a monkey (or human) without a visual cortex was blind. However, Humphrey (1970) argued that a monkey without visual cortex should have at least as much visual ability as a frog, since the role of tectum in directing whole body movements in frog (Lettvin *et al.*, 1959) is analogous to the role of superior colliculus (a mammalian midbrain [sub-cortical] homologue of tectum) in directing orienting movements in mammals. After two years of training, a monkey without visual cortex was able to use visual cues to grab at moving objects, and to use changes in luminance — such as an open door or obstacles with high luminance contrast from the background — for navigation, even though delicate processes of pattern recognition were never regained. Moreover, it was discovered that humans without visual cortex could also "see" in this action-oriented sense — but, remarkably, they were not conscious of the visual stimuli for their actions (Weiskrantz, 1974). As Stoerig and Cowey (1997) note: In human and monkey, destruction of the optic radiations and/or the primary visual cortex is characterized by an absence of any conscious vision, yet visual stimuli may nonetheless be processed. This unconscious vision includes neuroendocrine, reflexive, indirect, and forced-choice responses. In a provocative paper, Merker (2007a, 2007b) argues for "consciousness without a cerebral cortex," with a significant component of the neural underpinnings of consciousness located subcortically. However, the two statements are not equivalent. It is clear that cortical and subcortical structures work in concert in most behaviors, and that various sensory and motor functions survive limited cortical damage while others (dependent on the size and location of the lesion) are lost. Decerebration can still leave various forms of reactivity available, but there is no evidence that consciousness in the sense of concern here can survive this loss.

In Chapters 1 and 2 of this book, "Neurobiological Naturalism," and "The Evolutionary Origins of Consciousness," Todd Feinberg and Jon Mallatt focus on the evolution of "primary, phenomenal, sensory consciousness in vertebrate animals," offering a model based on neurobiological naturalism. They propose that such consciousness first appeared in the earliest vertebrates from mental reconstructions at the top of topographically organized neural hierarchies in the optic tectum of the midbrain (for vision and other senses) and in the pallium of the forebrain (for smell perception). They then argue that consciousness leaped forward independently in the first mammals and then in the first birds, to be generated by an expanding dorsal pallium (cerebral cortex). However, Humphrey's data on "What the Frog's Eye Tells the Monkey's Brain" and the data on blindsight in human and monkey leads me to take a different focus on consciousness (I will be more specific below) than any notion of consciousness that could be ascribed to an "early vertebrate," let alone a modern frog. Note, too, that a person with blindsight is indeed conscious — it is just that they have no conscious access to visual perception. The point, though, is that being able to react to stimuli in the environment must be distinguished from being conscious of such stimuli.

Further clinical studies of human patients with intact visual cortex showed a double dissociation between the "declarative" ability to communicate the size of an object, whether verbally or by pantomime, and the "procedural" ability to act upon objects. Goodale *et al*. (1991) studied a patient (DF) who had developed a profound visual form agnosia following a bilateral lesion of the so-called ventral pathway from primary visual cortex to temporal cortex, even though the dorsal pathway from occipital lobe toward the parietal lobe appeared to be intact. When the patient was asked to indicate the width of any one of a set of blocks either verbally or by means of her index finger and thumb, her response bore no relationship to the dimensions of the object and showed considerable trial to trial variability. Yet, when she was asked simply to reach out and pick up the block, the peak aperture between her index finger and thumb (prior to contact with the object) changed systematically with the width of the object, as in normal controls. A similar dissociation was seen in her responses to the orientation of stimuli. In other words, DF could preshape her hand accurately, even though she appeared to have no conscious appreciation (either verbal or by pantomime) of the visual parameters that guided the preshape.

Jeannerod *et al*. (1994) reported the case of a woman (AT) who was the "opposite" of DF. AT had a lesion of the occipito-parietal region that interrupted the dorsal route of visual processing, but left the inferotemporal lobe and the ventral pathway intact. AT could verbalize the diameter of a cylinder and could use her hand to pantomime its size, but could not preshape appropriately when asked to grasp it. Instead of an adaptive preshape, she would open her hand to its fullest, and only began to close her hand when the cylinder hit the "web" between index finger and thumb. But there was a surprise! When the stimulus used for the grasp was not a cylinder (for which the "semantics" contains no information about expected size), but rather a familiar object — such as a reel of thread, or a lipstick — for which the "usual" size is part of the subject's knowledge, AT showed a relatively adaptive preshape. This suggests that the ventral pathway provides the parietal areas with "default values" of action-related parameters for such objects, i.e., values which can serve in place of actual sensory data to, e.g., represent the approximate size of a known object to help the parieto-frontal system.

We thus see that consciousness is not a direct property of having neurons of a particular structure or complexity, because the visual data can be represented in two networks — the dorsal and ventral visual pathways — of comparable neural complexity, yet be accessible to consciousness only when the ventral network is intact. On this account, it is not the details of morphology or neurochemistry that distinguish the neurons of a "conscious" from an "unconscious" mode of operation so much as the larger circuits and systems of which they are a part. Some might argue that the blindsight patient or DF is perfectly conscious of the visual cues they use to guide action, but either cannot or will not give verbal or pantomimed expression to their consciousness, but I discount this possibility and suggest that the above examples support a critique of some of the other views of consciousness presented in this volume. Specifically, since there are no data whatsoever suggesting that the damaged portions of DF's brain contain microtubules different from those of the regions that support successful "non-declarative" hand movements, there seems to be no evidence for the role of microtubules in consciousness espoused by Hameroff and Penrose in this volume. In a related review paper, Hameroff and Penrose (2014) recall their proposal from the 1990's that consciousness depends on biologically orchestrated coherent quantum processes in collections of microtubules within brain neurons. Dramatically, they claim that their theory "establishes a connection between the brain's biomolecular processes and the basic structure of the universe" and that "consciousness plays an intrinsic role in the universe," but it is clear that their notion of consciousness is very far removed from that espoused by Feinberg and Mallatt, let alone the effort here to understand the sense in which human consciousness is far removed from that (if any) of the frog.

Natural selection can operate on the macromolecular building blocks of cells, on crucial cellular subsystems, and on the morphology of cells themselves, as well as the connectivity of these cells and their formation into diverse nuclei. What is selected about a subsystem, then, may be the impact of some change on a larger system or on a smaller detail, rather than the immediate change in the subsystem itself. Moreover, we understand that the genetic code may not specify adult forms so much as the processes of self-organization in cell-assemblies which can yield "normal" connectivity for the adult raised in a normal environment. Further, the environment which fosters adaptive self-organization may be as much social as physical in nature.

The manifold details here are beyond the scope of this chapter. Rather, my aim in this chapter is not to seek a magic transition from totally non-conscious other species to conscious humans, but rather to analyze how differing styles of behavior and social interaction may support an evolutionary path which renders plausible the emergence of consciousness in its human form. Given this, I consider consciousness to range from an almost reflex awareness of basic motor, sensory (which may include awareness of subtle social cues), and motivational states (so that even "simple" mammals may be aware of the difference between feeling maternal and feeling enraged) up to the subtle human consciousness that slips in and out of elaborate verbal arguments. I will give particular attention to how human development depended on diverse forms of language and communication, thus leaving room to appreciate the role of culture and bio-cultural events in the development of human consciousness.

3.3. Brain States, Mental States, and Consciousness

To move forward, we need to clarify the distinction between brain states and mental states. A human brain may be characterized in terms of hundreds of regions which may be distinguished, for example, by the types of neurons they contain, the chemical properties of those neurons, the arrangement of the neurons into layers or other structures, as well as by the patterns of connections they may have with receptors, effectors, and other regions of the brain. Each region in turn may contain millions, or even hundreds of millions of neurons. Current models of the brain focus, in particular, on two properties of neurons in a network:

(i) the "output" of each neuron (the pattern of electric signaling on its *axon*, the output fiber that may branch again and again to connect it to hundreds or thousands of other neurons).

(ii) the "strength" or "weight" (plus or minus, excitatory or inhibitory) of each *synapse*, i.e., each point where the axon of one neuron impinges upon another neuron. This *synaptic weight* measures the extent to which the signal in the first neuron's axon will affect the activity of the second neuron.

A great deal of work on learning and memory holds that the learning process changes the synaptic weights of the brain in such a way that on a later occasion the activity of the network will be changed to allow recall of a specific memory or exhibition of an acquired skill. The upshot of all this it that we can describe the state of the brain at two levels (and this is still only a partial approximation — for example, some synapses may play a neuromodulatory role, rather than being excitatory or inhibitory):

(i) *The activity state:* The current output of each individual neuron. The dynamic interaction of these "firing patterns" determine the brain's "computations" that underlie perception, thought and action, including those processes that in humans involve language.

(ii) *The synaptic state:* The current synaptic weight of each synapse. These "connection strengths" crucially determine the dynamic fate of a specific firing pattern — the same activity which might have faded without trace yesterday, might today trigger a strong emotional interaction because some intervening traumatic event has been encoded in the brain's synapses.

These states are of overwhelming complexity, with perhaps 10^{11} neurons and 10^{15} synapses in a single human brain. How do these then relate to mental state? Put simply, the mental state is a gross abstraction of the brain state which is at *approximately* the same level as language. The locution "at approximately the same level as language" is meant to convey two things: (i) that a mental state may *or may not* be expressible in words, while (ii) asserting that the level of detail that we ascribe to a mental state is far cruder than a brain state.

For example, I may see a face without being able to place it. The words "I've seen that person before — but where?" do not exhaust the mental state, for the latter includes the experience of facial features that we cannot put into words. If and when I recall the face, "Oh yes, during intermission at the opera last Saturday", the transition may be inexplicable at the "mental level", involving subtle neural processes which retrieved a memory of the scene involving that face, a representation rich enough to ground recognition of the context, with the subsequent return to the "mental level". The point here is that brain states are immensely complicated and causally complete. By this I mean that, brain states form a dynamic system: current state together with input determines (albeit with, perhaps, some noise) the next state and the output. By contrast, mental states are relatively simple and thus only sometimes causally efficacious — details of the brain state that we do not think about may be more important in determining our next thought than the preceding thought itself. In this regard, logic is not the essence of the mental, but is rather a crystallization of the limiting case where decisions and inferences can be made via inferential chains at the mental level with limited recourse to the brain's greater patternings.

Again, if we admire a sunset, our mental state involves our awareness and aesthetic appreciation of the rich patterning of red, orange and purple in the cloud formations banked above the horizon — the words "Look at that!" may then suffice to help a friend share aspects of that mental state without in any way reducing that state to those three neutral words. But the extent of that aesthetic appreciation does not begin to exhaust the complexity of the brain states which flash through the brain millisecond by millisecond as we enjoy the scene.

The relation of words to the above two vignettes is interesting. In the first case, the words are parts of our own thoughts, not tools for communication, and act as a précis of the current mental state that is itself a précis of the much more complex (and temporally extended) brain state. In the second case, the words we use for communication have nothing to do with the scene — but they convey the expectation that the friend too will appreciate the view of that sunset, an expectation that may be based either on one's knowledge of the friend's aesthetic preferences or on two more general propositions: "Most people enjoy watching a good sunset", and "It's good to increase the enjoyment of others." However, I again emphasize that logic is not "the essence of the mental," and it would in general not be the case that any explicit inference based on these two propositions was involved in your coming to say "Look at that!" I reiterate that causality is at the level of brain states, not mental states.

It should also be stressed that much of the brain state is irrelevant to one's mental state. For example, the state of the neural networks involved in the regulation of breathing do not affect our mental state save in times of crisis, and even then the effect on our mental state may be based on awareness of our overt pattern of breathing, rather than based on any monitoring of the relevant neural activity. But beware of that word "monitoring." We are *not* offering a dualist theory in which mind and brain are separate, with the mind monitoring the activity of the brain to extract highlights. Rather, we hold that the mental state is itself captured within the activity of the brain, forming some sort of précis of the broader neural activity *and* memory structures.

3.4. Consciousness, Language, and Distributed Schemas

Butler and Hodos (4) show that the course of brain evolution among vertebrates has been determined in great part by

(a) Formation of multiple new nuclei through elaboration or duplication;
(b) Regionally specific increases in cell proliferation in different parts of the brain; and
(c) Gain of some new connections and loss of some established connections.

These changes can be accompanied by exquisite variations in neural morphology. These phenomena can be influenced by relatively simple mutational events that can thus become established in a population as the result of random variation. Selective pressures determine whether the behavioral phenotypes expressive of the central nervous system organization produced by these random mutations increase their proportional representation within the population and eventually become normal for a new species. What needs stressing here is that the genome does not come neatly packaged in terms of separate sets of genes for separate nuclei of the brain, nor does each nucleus control its own set of behaviors. At the lowest levels, the specification of a brain region can call on generic hierarchies of gene regulation and expression for the formation of boundaries amongst generic precursors of, e.g., neurons and glia. At a higher level, late structural genes that typify a taxon may elaborate on these in multifarious and marvelous ways. At the highest levels, an overt behavior may reflect a) the interactions of multiple brain regions and b) the social interaction of creatures with each other and their environment.

Moreover, and of great relevance to our later analysis of the evolution of language, *behavioral phenotypes* are not the result of "brain genes" alone; rather, they express both the brain's inherent organization, and the learning that has shaped it through the experiences of the individual organism, and these are determined in great part by the social milieu in which the organism is raised. For many species, this "social milieu" is hard to disentangle from the biology, but for primates we can discern a variety of "rituals", "practices" and "tribal customs" which constitute a body of culture constrained by, but in no sense determined by, the biological make-up of the social group. Thus, as we come to analyze the evolution of the hominids, culture comes to play an important role even in biological evolution, as well as being itself subject to change and selection.

Here, niche construction plays a crucial role (Arbib & Iriki, 2013; Iriki & Taoka, 2012). *Niche construction* refers to the process whereby creatures, by changing the physical environment, in turn change the adaptive pressures that will constrain the evolution of their own and other species (Odling-Smee *et al.*, 2003). What we add here is that by creating new patterns of behavior, creatures can alter the *cultural* niche in which they evolve (Arbib, 2011; Bickerton, 2009; Pinker, 2010). Importantly, such cultural niche construction can take place on both the biological time scale and on the much faster time scale on which culture evolves, with an increasing tempo in the history of *Homo sapiens*. Indeed, humans can induce such changes intentionally, to create a novel environmental niche, thus *intentional* niche-construction (Iriki & Sakura, 2008). Going further, Iriki & Taoka (2012) advanced the idea of *neural* niche construction, the notion that new behaviors may remodel the brain in such a way that creatures can take advantages of opportunities that would not otherwise be possible, thus rendering genes adaptive that can modulate the brain around this new stable state.

It should also be noted that certain genes seem to control the coordinated development of "brain systems" involving many brain regions, rather than being "specialists" for some local neural structures. From the work of Holmes (1939), we view the cerebellum's role as crucial to the graceful adaptation and coordination of movements, while conceding that the plan of movement is elaborated elsewhere. In mammals, cerebral cortex and cerebellum seem to be coupled in a genetically controlled system in such a way that they can co-evolve so that it now appears that lateral cerebellum can contribute to even the highly cognitive aspects of neural function (Ito, 2012).

Without wishing to equate consciousness with motor control — far from it — I would suggest that consciousness, too, may be a system function that involves networks including but not necessarily limited to cerebral cortex, and that as cerebral cortex evolves, so does consciousness. But just as cerebellar adaptation for eye movements is a far cry from cerebellar adaptation for, say, piano playing or speech, so we may expect the quality of consciousness to vary from species to species.

We have argued that we should not seek an all-or-none transition from other species viewed as totally non-conscious to conscious humans, but rather seek an evolutionary path which renders plausible the emergence of consciousness in its human form. I argue that we are conscious in a fully human sense only because we have language — i.e., that as awareness piggy-backs on all manner of neural functions, so too must it piggy-back on language, thus reaching a subtlety and complexity that would otherwise be impossible. However, I strongly deny that consciousness is merely a function of language. For example, as already noted in the previous section, one can be aware of the shape and shading and coloration of a face in great subtlety, and be totally unable to put one's vivid, conscious perception of that face into words. This point is well illustrated by Georgia O'Keefe's statement, painted on the foyer wall of the Georgia O'Keefe Museum in Santa Fe, that "The meaning of a word to me is not as exact as the meaning of a color. Colors and shapes make a more definite statement than words. I am often amazed at the spoken and written words telling me what I have painted." Nathalie Sarraute, in her book *Childhood*, says

> For me ... there is something prior to language: a sensation, a perception, something in search of its language, which cannot exist without language … Scarcely does this formless [feeling], all timid, and trembling try to show its face than all powerful language, always ready to intervene so as to re-establish order — its own order — jumps on it and crushes it. Even the simple word "joy" cannot gather up what fills me, brims over in me, disperses, dissolves, melts into the pink bricks, the blossom-covered espaliers, the lawn, the pink and white petals, the air vibrating with barely perceptible tremors, with waves…waves of life, quite simply of life, what other word?…of life in its pure state… (Sarraute, 1984).

What are we to make of that phrase, "which cannot exist without language?" Sarraute sees language as both expressive and destructive, and that tension between the verbal and non-verbal is surely the hallmark of our consciousness, setting it apart from whatever form of consciousness may be experienced by other creatures. The problem is that the writer, and of course the speaker, must try to express within words that which goes beyond words. I shall return to this topic in the section on motivation and emotion, offering some preliminary thoughts on the role of emotions and feelings in the development of consciousness.

Elsewhere (Arbib, 1978, 1992, 2013b), I have developed a theory of schemas as functional, as distinct from structural, units in a hierarchical analysis of the brain. The starting point was to describe perceptual structures and distributed motor control in terms of functional units called schemas which may be combined to form new schemas as coordinated control programs linking simpler (perceptual and motor) schemas, but these schemas provide the basis for more abstract schemas which underlie thought and language more generally. Thinking in terms of schemas reminds us that the behavioral phenotype of an organism is not necessarily linked to a localized structure of the brain, but may involve subtle patterns of "cooperative computation" between brain regions which support schemas in a distributed fashion. Selection may thus act as much on schemas and hence through neural systems as it does on localized neural structures. Developing this view, Arbib and Liaw (1995) argue that evolution not only yields new *schemas* connected to the old, but yields reciprocal connections which modify those older schemas. Three principles summarize and extend the key points of the above discussion:

Principle 1: Cooperative computation. The functions of perceptual-motor behavior and intelligent action of animals situated in the world can be expressed as a network of interacting schemas. Each schema itself involves the integrated activity of multiple brain regions. The method of interaction of schemas is "cooperative computation" (competition/cooperation) so that "computations" which are often seen as the province of traditional symbol-based processing are carried out by distributed "neuron-like" methods which do not involve explicit symbolic control. Cooperative computation not only serves as a basis for coordinated motor actions, but even for reactive planning, and intelligent behavior, including the use of language.

Principle 2: Evolution and modulation. New schemas and brain regions often arise as "modulators" of existing schemas or brain regions, rather than as new systems with independent functional roles.

Principle 3: Interaction of partial representations. A multiplicity of different representations — whether they be partial representations on a retinotopic basis, abstract representations of knowledge about types of object in the world, or more abstract "planning spaces" — must be linked into an integrated whole. Such linkage, however, may be mediated by distributed processes of competition and cooperation. There is no one place in the brain where an integrated representation of space plays the sole executive role in linking perception of the current environment to action.

In the next section, we will see the way in which Principle 2 was anticipated by the work of 19th century neurologist Hughlings Jackson, and suggest that this may offer a crucial clue as to the specific character of human consciousness. In the process, we may extend the observation of the last sentence of Principle 3 to state: "*There is no one place in the brain where an integrated representation of experience provides the essence of consciousness*."

3.5. From Social Cooperation to Consciousness

Primitive communication subserves primitive coordination of the members of a social group, but processes which coordinate a group need not involve consciousness. Indeed, the social insects demonstrate a subtlety of group coordination that is in no way a precursor to consciousness. Here, we present the Arbib & Hesse (A & H) scenario for how group coordination may indeed have laid the basis for the co-evolution of consciousness and language (Arbib, 1985; Arbib & Hesse, 1986).

We each have only one body to act with and thus have a limited set of actions available to us at any one time. Thus, as we move towards the actual commitment of the organism to action, there would have to be a channeling from the richness of mental (and *a fortiori* neural) states to the well-focused choice, not necessarily conscious, of a course of action. Large ensembles of schemas in some sense interact, compete, cooperate to constitute a relatively well-focused plan of action that will commit the organism. This combination of many schemas within one body suggests a continuity of behavior by the one individual in similar situations, but also, as this repertoire builds up over time, the possibility that the schemas may eventually cohere in new ways, so that what had been an expected behavior in a certain set of situations may eventually give way, through new patterns of schema interaction, to new courses of behavior. Each individual has sets of schemas with some sort of coherence between them (this is not to claim that all of an individual's schemes are coherent); and the style of such a set of schemas can change over time to provide some sort of unity. I speculate that there are perhaps hundreds of thousands of schemas corresponding to the totality of what a human knows, with perhaps hundreds of these active at any time to subserve the current interaction of the organism with its environment. But what is the role of consciousness here? Would not these schemas do their jobs just as well if there were no such thing as consciousness? I cannot make a case for why we must *be* conscious. I think I can make a case for why we have patterns of schema activity that *correlate* with consciousness very well. But first, I do want to stress that our common sense about consciousness can be misleading, that there are many things that we think require our conscious thought but which do not — as we saw in our discussion of blindsight and DF.

The lesson is that even motor behaviors (reaching for a target, avoiding obstacles) that we think of as normally under conscious control can in fact proceed without our being conscious of the visual cues involved. A general point here is that what we may capture in correlates of a phenomenon within a limited account of one brain region pales before the richness of multi-level interactions playing out over tens, possibly hundreds of brain regions. The relation of visual cortex to superior colliculus in our exposition of blindsight can be related to the view of the British 19th century neurologist Hughlings Jackson of the brain in terms of levels of increasing evolutionary complexity (Jackson, 1874, 1878–1879). Jackson was very much influenced, as many people were at that time, by Darwinian concepts of evolution. In many cases, Jackson saw damage to the brain not so much in terms of the loss of function of the damaged region but rather in terms of the release of "older" brain regions from inhibition by the damaged area, which provided controls evolved later. In other words, such brain damage could often be seen as revealing behavior that is more primitive in evolutionary terms. As the brain evolves, new patterns of neural activity provide new "ecological niches" for the evolution of neural circuitry to exploit these patterns. Once such new circuitry evolves, there is a new "information environment" for earlier circuitry, which may thus evolve in turn to exploit the new patterns. (For an example of such a Jacksonian mode, see the account of optic flow in Arbib (1989a, pp. 342–344).)

Not only does the brain so evolve that new brain regions or schemas can exploit the prior richness of the brain but moreover — and this is the important point — *once these new systems are available they provide a richer environment for the older ones* which now have new possibilities for further evolution, whether evolution of brain regions over a biological time scale or the evolution of schemas over an individual time scale. If we compare, for example, the brain of a frog with the brain of a cat, we find that the corresponding regions of the visual mid-brain are much richer in the cat, not so much in terms of retinal input, but because there are so many descending pathways whereby the richness of cerebral cortex can influence what goes on in these more classical brain regions. In the terminology of schema theory, *evolution not only yields new schemas connected to the old, but yields reciprocal connections which modify those older schemas*.

Perhaps the development of animal and human communication may also be seen in terms of such evolutionary interaction to give us some insight into how consciousness might have evolved. A striking feature of human ability is that we come to incorporate tools into our body schema (Head & Holmes, 1911; Maravita & Iriki, 2004). Analogously, as creatures developed as social animals, their schemas came to represent their "bodies" as ending not at the extremities of the physical body but rather as extending to incorporate aspects of other members of the group. When we use a screwdriver our body ends at the tip of the screwdriver, not at the end of the hand; when we drive a car, our body ends at the rear bumper, not at our buttocks. Analogously, as creatures developed as social animals (and this account is not restricted to humans), the body might end not at the extremities of the physical body, but extend to incorporate aspects of other members of the group. However, coordinating others is a more subtle matter than just directing an arm or slightly adapting the hand to control a tool. The social animal has to find a way of expressing some précis of its mental state, of its richness of schema activity, so that it may then impinge upon others so that their behavior may be coordinated.

The A&H hypothesis of the evolution of human consciousness (Arbib & Hesse, 1986; Arbib, 1985) is that, with increasing richness of social interactions, though still at a prelinguistic stage in our evolutionary story, there would come the ability to form a *précis of schema activity* that is not necessarily relevant to deciding what to do next, but is relevant in terms of coordination with others. It is important to note that there is no claim that, at the earliest stages of its evolution, such a précis plays any role in deciding what to do next. However, in our "Jacksonian" terms, we can see that once the précis is available as part of the way the brain has evolved, older patterns of schema interaction may become modified in turn. There are occasions on which they can be better coordinated via the précis, and thus it would come about that the précis would develop into diverse states of a subnetwork (not necessarily localized) with bidirectional links into the schema network, and that this subnetwork would then evolve to have a role that is *sometimes* directive.

Primitive communication subserves primitive coordination of the members of a social group. As in blindsight, processes which coordinate group members need not involve consciousness. As communication evolves (by mechanisms we only begin to understand), the "instructions" that can be given to other members of the group increase in subtlety. Communication evolves at first purely as a way of coordinating the actions of a group. For this to succeed, the brain of each group member must be able not only to generate such signals, but also to integrate signals from other members of the group into its own ongoing motor planning. We suggest that the key transition in going from the limited set of vocalizations used in communication by, say, our common ancestors with vervet monkeys to the richness of human language came with a *migration in time* from:

(i) An execution/observation matching system enabling an individual to recognize the action (as distinct from the mere movement) that another individual is making, to
(ii) The individual becoming able to pantomime "this is the action I am about to take".

In the earliest stages of the evolution of this second ability, communication may have involved the accidental release of a motor plan from inhibition, thus allowing a brief prefix of the movement to be exhibited before the full action was released — but this "warning gesture" may have sufficed to alert others in time to bias their action, yielding benefits of adaptive value for groups that could both offer "signals of intention" and make use of them. This yields positive reinforcement to the individual accidentally releasing prefixes of actions, serving in turn as the basis for group selection, favoring the reproduction of those groups which can learn to emit and interpret such signals. This marks, at both the individual and species levels, the beginning of real communication as distinct from the release of signals. Note that we are not positing a form of Jungian *group consciousness* but only that *group selection* yields a population equipped to communicate more effectively and thus come to develop culture and cultural evolution which supports the sharing of ideas and ideologies in a manner that might be interpreted as group consciousness but which I would rather see as establishing memes and social schemas whose content is elaborated far beyond what is encoded genetically, which includes basic aspects of the motivational system.

However, Arbib and Hesse do not emphasize this external process of "group selection" in the population as a whole, but rather the changes within the individual brain made possible by the availability of a "précis" — a gesturable representation — of intended future movements (as distinct from current movements). They use the term *communication plexus* for the circuits involved in generating this representation. The Jacksonian element of their analysis is that the evolution of the communication plexus provides an environment for the further evolution of older systems. They suggest that once the brain has such a communication plexus, then a new process of evolution begins whereby the précis comes to serve not only as a basis for communication between the members of a group, but also as a resource for planning and coordination within the brain itself. This "communication plexus" thus evolves a crucial role in schema coordination. Elsewhere (Arbib, 2001), I asserted that "it is the activity of this co-evolved process that constitutes consciousness. As such it will progress in richness along with the increased richness of communication that culminates as language in the human line." I need to qualify that. I would rather say that "it is the activity of this co-evolved process that endows human consciousness with properties denied to other species," and add that the evolution of this novel brain system makes possible (Jackson again) the evolutionary enrichment of other systems in a way that makes it possible to enrich "animal awareness" with properties that other animals lack. What needs to be emphasized is the essentially distributed nature of the brain's representations (recall Principle 3 above). An anecdote may help with this. In 1961, when I was a novice PhD student at MIT, the eminent visual neurophysiologist Horace Barlow was discussing with me the research of Hubel and Wiesel (1959) who had discovered that cells in primary visual cortex of cats responded best to oriented edges crossing their receptive fields. Positing that such local data could be integrated to obtain a neural representation of distinctive contours in the visual field, Barlow said "And that's how you can recognize people from caricatures" to which I responded "But then how do I know you are not a caricature." This exchange left me with the understanding that any limited representation (such as that conveyed by words or communicative gestures) remains a pale approximation (caricature!) of what we can consciously experience. This reinforces the point: activity of the "communication complex" is not even approximately the extent of consciousness unless it is integrated into the activity of diverse systems (coordinating diverse schemas at multiple hierarchical levels) — just as activity in the visual system which provides contours that may aid the recognition of a familiar face in no way exhausts what we may experience on viewing that face in an image or when meeting the person in real life. We have already noted that a mental state is far more crude than a brain state is and, of course, the description of mental states by folk psychology's linguistic labels is even more limited. In the same spirit, we may note Elman's (2009) and Elman *et al.*'s (2004) discussion of words as "cues" for diverse mental entities, since a given word may have different meaning in different contexts and the earlier discussion of "Language, Metaphor and a New Epistemology" by Arbib and Hesse (1986) in which meaning emerges through the spread of activity in schema networks.

There are occasions on which "older" schemas can be better coordinated if they coordinate via the précis than if they coordinate by themselves. We then have a subnetwork of the schema network which provides a précis that may often have no directive role, and yet may evolve to have a role that is sometimes but not always directive. Since lower-level schema activity can often proceed successfully without this highest-level coordination, consciousness may sometimes be active, if active at all, as a monitor rather than as a director of action. In other cases, the précis of schema activity plays the crucial role in determining the future course of schema activity, and thus of action.

This evolutionary process, which occurs with subhuman species, then sets the stage, I would suggest, both for human consciousness and also for the evolution of language to express this rather coarse précis of the richness of the underlying schema activity. In summary, this theory sees human consciousness enriched by a précis of schema activity, evolving in such a way that it can elaborate certain mental processes at the level of language and logic and is related in part, but not entirely, to communication.

Arbib and Hesse's theory does not explain the phenomenology of consciousness — i.e., the way consciousness "feels" to each of us — but it does accord well with this phenomenology. The fact that lower-level schema activity can often proceed successfully without the high-level coordination afforded by the communication plexus explains why consciousness may sometimes be active as a monitor rather than as a director of action. In other cases, the formation of the précis of schema activity plays the crucial role in determining the future course of schema activity, and thus of action — and this accords with those occasions in which we experience a conscious effort in weighing a number of courses of action before we commit ourselves to behave in a specific way. O'Regan (2011, 2012) and O'Regan and Block (2012) develop a sensorimotor account of consciousness that is quite consistent with, and in some ways complementary to, the approach developed here. Of particular interest may be his paper "How to Build a Robot that is Conscious and Feels" (O'Regan, 2012), since this makes explicit the hypothesis that consciousness is to be understood at larger scales of organization (perhaps those addressed by schema theory) that can be shared by brains and machines, rather than depending on aspects specific to subneural physics.

So far, little has been said of emotions and feelings. Here, we may recall *The Feeling of What Happens* and *Self Comes to Mind* where Damasio (1999, 2012) claims that the first step in generating a "protoself" comes from primordial feelings, which humans share with other species. Similarly, Panksepp (1998) argues that primitive emotional feelings appear to lie at the core of our beings, and the neural mechanisms that generate such states may constitute an essential foundation for the evolution of higher, more rational, forms of consciousness. He further argues that a neural understanding of "the self" may arise from a study of how various biological value-coding systems (emotional circuits) converge and interact with representations of the body. I will devote a section to "Motivation and Emotion" below. Although I will not assess Damasio's and Panksepp's ideas explicitly I will offer at least a hint of how emotions and feelings relate to the A&H theory of the evolution of consciousness. The general thesis remains: just as pain, fear, hunger and so on may be a part of "animal awareness," so will they be part of human consciousness, but enriched by the interaction of primordial systems with the communication plexus — an interaction which has changed and enriched these systems across the course of primate evolution, building atop prior changes through mammalian evolution more generally. However, I first devote a section to the mirror system hypothesis (MSH) of the evolution of the human language-ready brain, a section which concludes with a very brief discussion of the notion of "theory of mind," which is not a scientific theory but rather the general understanding one has of the minds of oneself and others.

3.6. The Mirror System Approach to Language Evolution

More than a decade after working with Mary Hesse on the theory of the evolution of consciousness (the A&H theory) presented in the previous section, I developed with Giacomo Rizzolatti a preliminary version of MSH of the evolution of the language-ready brain. MSH fills in some of the details missing in the A&H theory as this section will now outline. But first I need to summarize the basic idea of "mirror neurons," as first discovered by Rizzolatti's group in the 1990s.

Premotor cortex of the macaque monkey contains an area F5 in which are located neurons specific for control of manual actions. Significantly, F5 contains a subclass of such neurons that discharge not only when the monkey grasps or manipulates objects, but also when the *monkey observes the experimenter* perform an action similar to the one that, when actively performed by the monkey, involves activity of the neuron (di Pellegrino *et al.*, 1992; Rizzolatti *et al.*, 1996). We distinguish such *mirror neurons*, which are active both when the monkey performs certain actions and when the monkey observes them performed by others, from *canonical neurons* which are active when the monkey performs certain actions but not when the monkey observes actions performed by others. We say that area F5 is endowed with an *observation/execution matching system*: When the monkey observes a motor act that resembles one in its movement repertoire, a neural code for this action is automatically retrieved. This code consists in the activation of a subset, the mirror neurons, of the F5 neurons which discharge when the observed act is executed by the monkey itself.

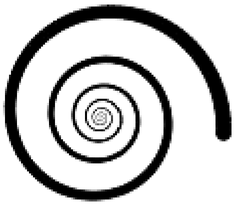
Two PET experiments (Grafton *et al.*, 1996; Rizzolatti *et al.*, 1996) were then designed to seek mirror systems in humans. The two experiments differed in many aspects, but both had a condition in which subjects observed the experimenter grasping a 3D object. Observation of the object alone was used as a control. (This condition also controlled for verbalization.) Grasp observation significantly activated the superior temporal sulcus (STS), the inferior parietal lobule, and the inferior frontal gyrus (area 45). All activations were in the left hemisphere. The last area is of special interest — areas 44 and 45 in left hemisphere of the human constitute Broca's area, a major component of the human brain's language mechanisms. Indeed, F5 is generally considered to be the homologue of Broca's area. But why should a mirror system for grasping be related to a system in the human brain for language production? For want of better data, we will assume that the common ancestor of humans and monkeys shared with monkeys a primate call system (a limited set of species-specific calls) and an oro-facial gesture system (a limited set of gestures expressive of emotion and related social indicators) — communication is inherently multi-modal, and body posture also plays a role in social communication, though I shall not emphasize this here. What is to be stressed here is that combinatorial properties for the openness of communication are virtually absent in basic primate calls and oro-facial communication, even though individual calls may be graded. Moreover, the neural substrate for primate calls is in a region of cingulate cortex distinct from F5, which we have seen to be the monkey homologue of human Broca's area. The *mirror system hypothesis for the evolution of the human language-ready brain* (Arbib, 2002, 2005b; Arbib & Rizzolatti, 1997; Rizzolatti & Arbib, 1998) suggests that human brain mechanisms supporting language evolved from a mirror system for manual actions in the last common ancestor of humans and macaques (LCA-m) in six stages:

Pre-Hominid

1. **A mirror system for grasping:** matching action observation and execution for grasping. Shared with common ancestor of human and monkey.
2. **A simple imitation system for grasping.** Shared with common ancestor of human and great apes.

Hominid Evolution

3. **A complex imitation system:** *complex imitation* combines the ability to recognize another's performance as a set of familiar movements with the ability to use this recognition to repeat the performance, and (more generally) to recognize that another's performance combines actions which can be imitated at least crudely by variants of actions already in the repertoire, with increasing practice yielding increasing skill.
4. **Protosign:** a manual-based communication system, breaking through the fixed repertoire of primate vocalizations to yield an open repertoire.



5. **Protospeech and multi-modal Protolanguage:** This rests on the "invasion" of the vocal apparatus by collaterals from communication system based on F5/Broca's area. Protosign and protospeech further co-evolve in an expanding spiral.

Cultural Evolution in Homo Sapiens

6. **Language:** The transition from action-object frames to verb-argument structures which express them; syntax and a compositional semantics: Co-evolution of cognitive & linguistic complexity.

In relation to Stages 4 and 5: Bickerton (1995) has spoken of *proto-language* as being language restricted, basically, to two word utterances (generally comprising a verb and a noun in some order), and suggests that it is common to chimps trained to exhibit some form of language-like use of symbols, two-year old humans, and people speaking pidgins. He then argues that proto-language was possessed by the hominid precursors of humans. However, I note that proto-language is only observed in creatures (chimps or humans) exposed to human language. Rather, I have argued that what early *Homo sapiens* possessed was the ability to name events with novel sequences of (manual and/or vocal) gestures, but that this capability does not imply the ability to separately name the objects and actions that comprised those events. I then claim that the latter ability was a momentous discovery made by humans perhaps 100,000 to 50,000 years ago, rather than a biological heritage from earlier hominids.

In relation to Stage 6: There is a danger, typified by Chomsky's notion of Universal Grammar (e.g., Chomsky, 1980), of thinking that there is some set of basic characteristics typical of all language and "hard-wired" into the human genome and thus the human infant's brain. I disagree, arguing that we can each know one or a few languages, but there is no such thing as Language-with-a-big-L that we all know, let alone "know in our genes". It is difficult to learn a foreign language because languages can be inherently different, but more-or-less adequate translations between languages are possible because languages have been shaped both by the need to express a range of basic human experiences, and by cultural diffusion.

The claim here is that biological evolution may not so much have yielded a human brain "equipped with" language so much as a human brain "ready for" language. The corollary to this is that even though an early *Homo sapiens* infant raised (after some miracle of time travel or cloning) in a human home in today's world would be completely human, an adult *Homo sapiens* of 200,000 years ago would have had a consciousness in many ways limited with respect to a modern human. Perhaps his awareness of the sights and sounds and smells around him would be more intense than most modern humans' experience, but it would lack the subtle overlays that the modern mind possesses precisely because of the recursive properties opened up by language.

Consciousness is shaped by the interaction of our biology (brain and body) with our environment (physical and social) and thus, I claim, has developed drastically since the human brain and body achieved their more-or-less modern genetic substrate. In this regard we may note the observation (Lee, 2000) that Whorf (1937) contrasts a "universe … [of] conscious thinking abstracted from experience" with a "universe as unconscious thinking projected upon experience" and refers to "the organization of raw experience into a consistent and readily communicable universe of ideas though the medium of linguistic patterns" and suggests that in referring to "conscious thinking" what Whorf had in mind was the elements of language that allow that experience to be organized and reflected upon in shareable form.

The MSH is by now an oft-told tale, and I will not repeat it in any detail here. A book length exposition has been published (Arbib, 2012) and a summary of the book (Arbib, 2013c) was published a year later along with 12 critiques and a response which carried the hypothesis forward (Arbib, 2013a).

But what is the relationship of the MSH of the late 1990s, whose development still continues, and the earlier A&H hypothesis that a précis of intended action could have provided the basis for the emergence of a uniquely human form of consciousness? Does the temporal sequence of the formulation and publication of these ideas have any relation to the evolutionary development of what they describe in primates? I would suggest that the answer is "no" — in other words, the evolution of the language-ready brain and of a specifically human consciousness are intertwined. The original A&H story has two components:

(i) The notion that having an expressible précis of intended action could improve social interaction.

(ii) The further notion that such a précis could then create a new resource for brain function, in the style of Hughlings Jackson, that could create a system whose activity could be seen as providing the correlate of specifically human dimensions of consciousness.

The MSH complements this by suggesting how neural mechanisms supporting the expression of that précis could have evolved in a manner that, crucially, can link the production and perception of such expressions. However, as Dominey (2013) observes, *How the Brain Got Language: The Mirror System Hypothesis* (Arbib, 2012) says rather little about "the uniquely human drive to share psychological states with others and unique underlying forms of cognitive representation for doing so, as illustrated in the research of Tomasello *et al*. (2005). This includes both the behavior itself, i.e., the motivation to share mental states, as well as the underlying neurophysiology." Item (ii) of A&H might offer a start in this direction: that what starts as a mechanism for coordinating behavior evolves linkages with diverse brain systems beyond the premotor cortex, and that these come to incorporate more details of processing related not only to current action but also to motivational states and perceptual states that are currently being experienced, whether or not they have immediate links to practical action. However, the mechanisms that may have first evolved to coordinate actions with others can now be exapted to express aspects of the précis as enlarged by these new connections, and thus the transition from instrumental to declarative communication begins. We can enrich this account further by a brief discussion of motivation and emotion, since another dimension of language is indeed the expression of emotion, whether through the words that are chosen or through prosody. It is worth reiterating that all this talk of a précis and of consciousness in no way implies the localization of consciousness in any one region of the brain. Indeed different lesions may affect different dimensions of consciousness, just as lesions can differentially affect the syntax, semantics and prosody of language. Ramachandran (2011) also articulates different levels of representations corresponding to different levels of consciousness, emphasizes the role of language in the emergence of consciousness, and sees consciousness as emerging from a number of interacting brain processes.

It has been argued that fully developed conscious experience requires some degree of self-awareness, and hence a *theory of mind*, the human ability to ascribe mental states to oneself and others. Much has been written about the possible role of mirror systems in the development of such theory of mind, with particular attention to its partial breakdown in autism (Arbib, 2007; Dapretto *et al*., 2006; Ramachandran & Oberman, 2006; Williams, 2008; Williams *et al*., 2001).

One referee observed of an earlier draft that "the claim that language decisively contributes to the emergence of fully developed human consciousness leads to the conclusion that consciousness is [in great part] *narrative* or *discursive*. … This is important in so far as it goes against a long tradition (especially in the continental philosophy) to speak of consciousness as a — more or less coherent — chain of mental images. Moreover, Arbib's theory seems also to undermine the idea of mentalese (language of thought). As I read Arbib's claims, he believes that there is more to consciousness than thought formulated in language (i.e., there are aspects of mental states that cannot be expressed in language), but at the same time he rejects the thesis that there exists a language of thought substantially different from the language we speak. These are just a couple of examples of philosophical doctrines which seem incompatible with Arbib's theory of consciousness; what they illustrate is that the conception proposed by Arbib is of a great significance for philosophy, as it is for cognitive science." Here I need to add two comments:

(a) The role of narrative in the evolution of language has been eloquently expressed by Barnard (2013) who argues that the evolution of narrative, and especially of myth, pushed the evolution of linguistic complexity, with language coevolving with mythology in symbolic frameworks which extended the capacity for verbal expression to the limits of cognition. My only demurrals are that narrative can perform valuable social functions without (or prior to) the emergence of myth, and that cognition co-evolved with means for verbal expression, rather than providing pre-existing limits.

(b) It is not that I "reject the thesis that there exists a language of thought substantially different from the language we speak" — I reject the notion that our mental processes are language-like save in limiting cases, with mental processing better defined at the level of schemas and as a mere abstraction of the even less language-like dynamics of brain states (Arbib, 1989b).

Given that the title of this chapter is "How Language Evolution Reshaped Human Consciousness," it may be surprising that my first paper on consciousness was entitled "Consciousness: The Secondary Role of Language" (Arbib, 1972). For our purposes, its key assertion was that "language arose as an elaboration of more fundamental functions of perception. This does not deny that the evolution of languages has greatly enriched human behavior, but does suggest the danger of trying to understand the human mind by regarding language as its primary attribute."

I suggested that the primary purpose of perceiving objects is not to name them, but rather to be prepared to interact with them — though you may perceive an object yet still leave it alone. In this spirit, the aim of the present chapter has been to suggest how the processes that marked the dramatic emergence of the human language-ready brain (or, more generally, the "culture-ready" brain (Whitehead, 2010) did not reduce human consciousness to thoughts expressible in language but, rather, have transformed more general forms of "animal awareness" to a uniquely human form of consciousness in which basic perception and awareness of action are enriched by a continuing "verbal commentary" on our awareness, and the ability to situate our current context beyond the here-and-now through what Suddendorf and Corballis (2007) refer to as "mental time travel."

3.7. Motivation and Emotion: The Motors of Behavior

We earlier quoted the emotive meaning of color to Georgia O'Keeffe and Nathalie Sarraute's description of her transcendent feelings which could not be reduced to language to make the point that our consciousness is by no means limited to what can be expressed in language. We experience nuances of awareness that go far beyond that which could be shared — let alone expressed — by an animal, and a skill of great writers is to find words that evoke these nuances of feeling and experience which most of us can seldom verbalize with any subtlety. In this section, I want to broach the issue of emotions and feelings in the development of consciousness. Simple mammals may be aware of the differences between "feeling maternal" and "feeling enraged," while the primate call system is a limited set of gestures expressive of emotionally charged behavior (e.g., fleeing a predator). However, MSH is at some pains to distinguish the medial brain systems implicated in primate calls from the lateral brain systems (like macaque F5) that seem homologous to the cortical areas most engaged in language processing.

This section is based on papers (Arbib, 2005a; Arbib & Fellous, 2004) that grew out of my work with Jean-Marc Fellous on the book *Who Needs Emotions: The Brain Meets the Robot*, (Fellous & Arbib, 2005), and was influenced by various contributions there. As suggested by the book's title, much of our concern there was in relating the neurobiology of animal and human emotion to the question of whether (future) robots could have emotions, and thus to the issue of what "emotion" would mean when cut loose from its biological moorings. Again we see the theme of trying to understand what aspects are specific to our biology, and which may in time be captured by advances in robotics and artificial intelligence, but here I will focus solely on the nature of human emotion and its roots in biological (and social) evolution. We analyzed emotion under two headings:

(1) *Emotional expression for communication and social coordination.*
(2) *Emotion for organization of behavior (action selection, attention, social coordination and learning).*

The former concerns "external" aspect of emotions; the latter "internal" aspects. In animals, these aspects have co-evolved. Animals need to survive and perform efficiently within their ecological niche and in each case patterns of coordination will greatly influence the suite of relevant emotions (if such are indeed needed) and the means whereby they are communicated. A key function of emotion is to communicate simplified but high impact information. A scream is extremely poor in information (saying nothing about the cause for alarm), but its impact on others is high. The similarity in facial expressions between certain animals and humans prompted classic evolutionary analyses (Darwin, 1872) which support the view that mammals (at least) have emotions, though not necessarily *human* emotions. Many researchers (e.g., Dalgleish, 2009; Damasio, 1994; Fellous & Ledoux, 2005; Jeannerod, 2005; LeDoux, 2012; Rolls, 2005) have proposed explicit functions for emotions: coordinating behavioral response systems; shifting behavioral hierarchies; communication and social bonding; short-cut cognitive processing; and facilitating storage and recall of memories. However, emotions are not always beneficial — if one is caught in a traffic jam, anger may easily set in, yet anger *in this case* has no apparent utility. How does the brain maximize the benefits of emotion while minimizing its occasional inappropriateness?

To echo something of the role of the précis in the A&H theory and the emergence of language in MSH, we now turn to general "cognitive architectures" in which the role of emotion can be situated at several levels. Ortony *et al.*, (2005) analyze the interplay of *affect* (value), *motivation* (action tendencies), *cognition* (meaning), and *behavior* at three levels of information processing:

Reactive: a hard-wired releaser of fixed action patterns and an interrupt generator. This level has only proto-affect.
Routine: the locus of unconscious well-learned automatized activity and primitive and unconscious emotions.

Reflective: the home of higher-order cognitive functions, including metacognition, consciousness, self-reflection, and full-fledged emotions.

They address the design of emotions in computational agents that must perform unanticipated tasks in unpredictable environments. They argue that such agents, if they are to function effectively, must be endowed with curiosity and expectations and a *sense of self* that reflects parameter settings that govern the agent's functioning. Sloman *et al*., (2005) also offer a three-level view of central processes:

Reactive: producing immediate actions. When inconsistent reactions are simultaneously activated one may be selected by a competitive mechanism.

Deliberative: using explicit hypothetical representations of alternative possible predictions or explanations, comparing them and selecting a preferred option.

Meta-management: allowing internal processes to be monitored, categorized, evaluated and controlled or modulated.

They further note the utility of an "alarm" system, a reactive component that gets inputs from and sends outputs to all the other components and detects situations where rapid global redirection of processing is required.

To reconcile the two schemes, Arbib and Fellous (2004) suggested four levels: reactive; routine; reflective-deliberative; and reflective-meta-management.

Emotion is not computed by a centralized neural system; emotions operate at many time scales and at many behavioral levels; and there is no easy separation between emotion and cognition. Each animal comes with a set of basic "drives" — for hunger, thirst, sex, self-preservation, etc. — and these provide the basic "motor", motivation, for behavior. Motivated behavior not only includes bodily behavior (as in feeding and fleeing, oro-facial responses, and defensive and mating activities), but also autonomic output (e.g., heart rate and blood pressure), and visceroendocrine output (e.g., adrenaline, release of sex hormones). These lie at the heart

[sic] of our emotional repertoire. However, the emotions that we talk about and perceive in others are both more restricted than this (how many people perceive another's cortisol level?) yet also more subtle, intertwining these basic motivations with our complex cognitions of social role and interactions, as in the cases of jealousy and pride. The core of the motivation system is provided by the nuclei of a deep-seated region of the brain called the *hypothalamus*. These nuclei are devoted to the elaboration and control of specific behaviors necessary for survival. Such behaviors include spontaneous locomotion, exploration, ingestive, defensive, and reproductive behaviors. Basically, the hypothalamus talks "downward" for basic behavioral control, and "upward" to involve the cortex in determining when particular behaviors are appropriate. The lateral hypothalamus plays a critical role in arousal, control of behavioral state, and reward-seeking behavior. It includes what Olds (1969) referred to as the "pleasure center" because rats will press a lever thousands of times per hour to deliver electrical stimulation to this region.

The role of the amygdala in the experience and expression of fear has received particular study. Stimuli that elicit fear reactions may be external (e.g., a loud noise) or internal, including those from memory structures such as hippocampus or prefrontal cortex. The amygdala can influence cortical areas via feedback from proprioceptive, visceral or hormonal signals, via projections to various "arousal" networks, and through interaction with the medial prefrontal cortex. Prefrontal cortex, in return, sends distinct connections to several regions of amygdala, allowing elaborate cognitive functions to regulate the amygdala's roles in emotion. Because of the tight interactions between amygdala and prefrontal cortex, it is likely that our ability to generalize and abstract is directed by (and influences, in return) some aspects of our emotional state. Connections between the amygdala, thalamus, and cortical areas allow for both fast emotion elicitation and more refined emotion control based on memory and high level sensory representations. This role of prefrontal cortex is exemplified by the observations that decision-making ability in emotional situations is also impaired in humans with damage to the medial prefrontal cortex, and abnormalities in prefrontal cortex may predispose people to develop fear and anxiety disorders (LeDoux, 2000). Moreover, activation of the human amygdala can be produced by observing facial expressions, and lesions of the human amygdala may cause difficulty in the identification of some such expressions (Rolls, 2000). The amygdala and prefrontal cortices are therefore involved in social as well as internal aspects of emotion, and together play a crucial role in the regulation of emotion, a key component of affective style, affective disorders, and social interactions (Adolphs, 2009).

The *medial prefrontal cortex* is adjacent to the *orbitofrontal cortex* and Rolls (2005) documents how damage to the orbitofrontal cortex produces emotional changes which includes a tendency to inappropriate responses. Orbitofrontal neurons also serve as part of a mechanism which evaluates whether a reward is expected, and different subregions of the prefrontal cortex are selectively involved during positive rewards or punishments (Schultz, 2000). Others have shown that different subregions of the medial temporal lobe memory system are selectively and differentially involved for emotional and neutral stimuli in human, and that this area was strongly correlated with amygdala activations during emotional stimuli. Many other brain areas have been involved in the experience and expression of emotions in humans, including the anterior cingulate cortex, insula, hippocampus, and fusiform gyrus.

LeDoux (2000) argues that emotion systems evolved as sensory-motor solutions to problems of survival. He distinguishes emotions from "feelings". Such "conscious emotions" are not, he suggests, the function that emotion systems evolved to perform. He has focused on fear-conditioning in the rat to define meaningful animal experiments and has been particularly concerned with the role of the amygdala in fear. In particular he has studied the role of the amygdala in conditioning of fearful behavior, as an animal comes to learn that certain situations may lead to danger. The laboratory equivalent might be for a rat to learn that it will get an electric shock if it approaches a particular place in its cage. LeDoux stresses that we cannot know what feelings the rat has if it is afraid, or indeed whether it has any feelings of fear. But it certainly can exhibit fearful behavior, trying to escape or, in many of LeDoux's studies, simply freezing in place. Such "freezing", a cessation of normal locomotion, is well known as a response to predators (if you do not move, he may not see you) and so is assumed to be a case of being too fearful to move.

Whether or not rats have emotional feelings as well as emotional behaviors, there is no doubt that humans do have feelings. How does "behavioral fear" relate to the feeling of fear or anxiety? The crucial element from an evolutionary point of view may be the set of reciprocal interactions between amygdala and cerebral cortex. Human emotions are strongly influenced by social milieu — what is embarrassing in one culture (e.g., nudity) may be regarded as perfectly normal in another. The amygdala can influence cortical areas by way of feedback either from proprioceptive or visceral signals or hormones, via projections to various "arousal" networks, and through the aforementioned interaction with medial prefrontal cortex. Fellous and LeDoux (2005) suggest that prefrontal-amygdala interactions may be involved in the conscious feelings of fear. However, this division between the cognitive cortex and emotional amygdala strikes me as too simple to do justice the brain's integration of widely distributed functions — both because not all parts of cortex give rise to conscious feelings and also because human emotions seems to be inextricably bound up with "cortical subtleties."

Going beyond the above hypothalamo–amygdala–cortical interactions, we note that language plays a unique role in shading and refining human emotions. It is thus interesting that recent research suggests that "mirror neurons" may provide a crucial part of the substrate for "empathy", the ability to recognize the emotional dispositions of others, as well as communication through language. We stress that the richness of human emotion is in part due to the linkage of emotion with cognitive processes of great subtlety. The neural substrate of emotion involves many interacting structures, from the hypothalamus to the orbitofrontal cortex to the diverse mirror systems, which have clear "cognitive" as well as emotional roles.

In tracing the macaque neurophysiology of the mirror system, we saw that neurons in the STS are active during movement observation though perhaps not during movement execution, and that STS and F5 may be indirectly connected via area PF which, like F5, contains mirror neurons. As Rizzolatti *et al*. (2001) observe, STS is also part of a circuit that includes the amygdala and the orbitofrontal cortex and so may be involved in the elaboration of affective aspects of social behavior. Human emotions are greatly influenced by our ability to empathize with the behavior of other people. We have all had moments when we felt a welling up of sadness on seeing the grief of another or smiled in response to another's smile. In *The Expression of the Emotions in Man and Animals*, Darwin (1872) observed that we share a range of physical expressions of emotion with many other animals. Our tendency to feel others' emotions, and the fact that certain emotions are so universal as to be shared with other mammals have led a number of researchers to suggest that in addition to a mirror system for manual actions, we have one for the generation and recognition of physical expressions of emotion. In fact, there seem to be multiple systems involving different classes of emotions.

The suggestion is that these systems enable us to empathize with others not simply by recognizing their emotions but by experiencing them. This is related to the *simulation theory* (Gallese & Goldman, 1998) which proposes that mirror neurons support our ability to read minds by "putting ourselves in the other's shoes", and using our own minds to simulate the mental processes that are likely to be operating in the other. Other findings suggest that the ability to recognize the auditory and visual details of a person's actions may be a factor in empathy (Gazzola *et al*., 2006). The role of a brain region called the insula in a mirror system for disgust (Wicker *et al*., 2003) suggests that emotional communication may involve diverse mirror systems. However, "emotion recognition" is not just a matter of empathy. If I were to see someone advancing toward me with a knife, I would recognize his anger yet feel fear; I would not empathize as I struggled to protect myself. And, indeed, much of the literature on mirror neurons does indeed distinguish between low-level visuomotor reaction and a more comprehensive understanding of the other's emotions. A crucial aspect of social interaction is our view of others. The term "Theory of Mind," briefly discussed at the end of Sec. 6, has become widely adopted for the ability of humans to understand that other people may be like themselves yet have different knowledge and points of view. Gallese and Goldman argue that observed action sequences are represented in the observer "off-line" to prevent automatic copying, as well as to facilitate further processing of this high-level social information. However, mirror neurons alone cannot mediate understanding. Instead, understanding rests on a network of associations distributed across multiple brain regions "beyond the mirror."

This cautions us to distrust talk of "the" mirror system but rather seek to understand the roles of, and interactions, between, multiple mirror systems such as those for hands and faces and — in humans — language. The previous section had its primary emphasis on intermediate stages in the evolution of mirror systems from the macaque-like mirror system for grasping to those that support language. I would suggest that, similarly, a number of stages would have to intervene in the evolution of the brain mechanisms that support emotion and empathy. This raises the challenge of connecting empathy with language evolution by assessing how our ability to experience others' emotions affected the way the language-ready brain evolved. But this challenge lies outside the scope of this chapter.

Appraisal theory (Ortony *et al*., 1988) develops a catalogue of human emotions and seeks to provide a computational account of the "appraisals" which lead to invocation of one emotion over another, but without the "heat" provided by emotion's biological underpinnings in humans.

Appraisal theory is thus a good candidate for algorithms that enable a robot to *simulate* the appearance of human emotional behavior. However, in biology, the 4Fs (feeding, fighting, fleeing, and reproduction) (Pribram, 1960) are paramount. S-RETIC (Kilmer *et al.*, 1969) modeled the brain's reticular formation as a stack of modules, each with a slightly different selection of input, but each trying to decide to which 4F-like mode to commit the organism. They would communicate back and forth, competing and cooperating until reaching a consensus that would switch the organism's mode. Within any mode of behavior, many different acts are possible: if the cat should flee, will it take the mouse first or leave it, climb a tree or skirt it? The notion is that a hierarchical structure that computes modes, and then acts within modes, might in some sense be "better" than one that tries to determine successive acts directly. Consider an animal with a set of basic functions, each with appropriate perceptual schemas and access to various motor schemas. Each perceptual schema evaluates the current state to come up with an "urgency level" for activating various motor schemas, as well as determining appropriate motor parameters. A competition mechanism can adjudicate between these processes. A *motivational system* is a state-evaluation process that can adjust the relative weighting of the different functions, raising the urgency level for one system while lowering the motivation system for others depending on the context. One may then generalize "modes" to abstract groups of tasks, with many strategies gathered into a small number of modes such that it is in general more efficient, when faced with a problem, to first select an appropriate mode and then select a strategy from within that mode. The catch is that there may be cases in which rapid commitment to one group of strategies may preclude finding the most appropriate strategy — possibly at times with disastrous consequences. The brain has to balance this "fast commitment" process against more subtle evaluative processes that can check the suitability of a chosen strategy before committing to it completely (and thus we return to the varied levels of "Cognitive Architectures"). We might then liken "motivation" to biases that favor one strategy group over another, and "emotion" to the way in which these biases interact with more subtle computations. One way emotions may become maladaptive, then, is when biases favoring rapid commitment to one mode overwhelm more cautious analysis of available strategies.

Recall our earlier introduction of two "views" of emotion: (1) emotional expression for communication and social coordination; and (2) emotion for the organization of behavior. The very utility of "modes" in decision making (sense 2 of emotion) may well make fast communication of modal commitment an effective way to coordinate animals or humans.

Emotion changes the operating characteristics of cognition and action selection. There is neuroscientific evidence that affect is a prerequisite for establishing long-term memories. It also has important consequences for the allocation of attention. Fear tends to focus attention on local details; but under conditions of positive affect people tend to focus on the bigger picture (Gasper & Clore, 2002). Resource mobilization and prioritization of behaviors rely on compact signals that have high-impact on the functioning of an autonomous agent. The underlying hypothesis is that hierarchical assessment of task priorities can help animals cope with the complexity and unreliability of the (external or internal) environment — mirroring the rich interaction of cognitive and emotional processing in the mammalian brain. Adaptive mechanisms can adjust the hierarchy and grouping of tasks to match the constraints imposed by time, physical limits, and energy resources. In this regard, note that human emotion becomes maladaptive not only when rapid commitment processes overwhelm more cautious analysis, but also when worry or other related emotions that favor more cautious analysis overwhelm adaptive rapidity.

3.8. Concluding Claims About Consciousness

To conclude briefly, the A&H theory of consciousness has the following ingredients.

(i) Human consciousness as we normally experience it is a property of the brain, rather than some separate "mind stuff".
(ii) Consciousness is a distributed property that has access to many brain regions, and provides a précis based on the state of other brain regions that is only partially verbalizable.
(iii) It is possible that portions of our brain can support forms of "animal awareness" that may enrich human consciousness but seem qualitatively different in nature; but
(iv) What makes human consciousness so different is that it includes expression of our thoughts in language.
(v) The *communication plexus* underlying language has (by a process of "Jacksonian evolution") restructured the brain in such a way that consciousness may seem to be sometimes observer and sometimes controller.

The third acknowledges a possible evolutionary continuity in consciousness while the fourth stresses that the consciousness of humans who possess language is qualitatively different from that of other creatures; while the last ascribes a specific role to the evolution of a "communication plexus" for the communication of intended action in the emergence of human consciousness.

MSH complements the A&H theory by offering a more "brain-based" approach to the evolution of human language, tracing it through seven stages: (1) grasping, (2) a mirror system for grasping, (3) a "simple" imitation system, (4) a "complex" imitation system, (5) a manual-based communication system, (6) speech, and (7) language. Stages 1 and 2 are present in the monkey and posited to be shared by the last common ancestor of monkey and human; stage 3 is posited to be more fully developed in the last common ancestor of human and chimpanzee; stages 4, 5 and 6 are posited to occur along the hominid line. However, we do not argue that the "protospeech" of the first *Homo sapiens* was a human language in anything like the modern sense. Rather, the development of human languages and the attendant enrichment of human consciousness was, I argue, a process that took many tens of millennia.

We also outlined an account of the roles of empathy and emotion in the emergence of human consciousness. In this regard, it is worth noting the complementary nature of language and music (Arbib, 2013d; Arbib & Iriki, 2013; Levinson, 2013; Seifert *et al*., 2013), the former more adapted to the expression of propositional content, the latter more directly linked to the evocation of emotion, suggesting that further understanding of their relationship will indeed enrich the approach to the emergence of consciousness sketched in these pages.

To conclude, I should mention that there is now a range of research linking theories of consciousness to data on the brain, but analysis of the recent material is outside the scope of this chapter. A fine example of such an approach is given by the review of "Experimental and Theoretical Approaches to Conscious Processing" by Stanislas Dehaene and Jean-Pierre Changeux (Dehaene & Changeux, 2011) which links data on macaque neurophysiology and human brain imaging with explicit modeling via computational neuroscience. However, it is worth noting that although the aim of their study is more ambitious in its attempt to review current neuroscience data it is also less ambitious in that it is limited in its delineation of consciousness: its main concern is with one question: "When a visual stimulus is presented to a human subject, they may or may not be able to declare (whether verbally, or by some motor response specified by the experimenter) what they have seen. What differences in neural activity characterize the difference between being aware vs. unaware of a visual stimulus?" The distinction is made more interesting by the fact that a visual stimulus may serve to prime a later response even though the subject remains unaware of the first stimulus.

In developing a theory to integrate the psychological and neuroscience data they review, Dehaene and Changeux situate themselves in agreement with a range of models of consciousness (see their paper for detailed citations) which emphasize four concepts:

- A supervision system,
- A serial processing system,
- A coherent assembly formed by re-entrant or top-down loops,
- A global workspace for information sharing. The theater metaphor

and indeed they call their framework the *global neuronal workspace* (GNW) hypothesis, and propose that associative perceptual, motor, attention, memory, and value areas interconnect to form a higher-level unified space where information is broadly shared and broadcast back to lower-level processors. My theory agrees to a limited extent (it sees supervision and serial processing as only partially in play, being side-effects of inherently distributed processes of competition and cooperation), but gives group coordination and the emergence of language (i.e., social rather than purely internal phenomena) a special evolutionary role. I suggest that language evolved for communication but as a corollary enriched our consciousness with an internal monolog. To simplify, I would claim that a frog has "no" consciousness (though it can adaptively react within its environment); a monkey has an extended animal awareness but no access to language; and a human has all this plus a "running commentary" that can structure or annotate experience but may only intermittently play a crucial role in committing the person to a course of action. Data are emerging that are relevant to all this, but their analysis to test my claims must await another day.

Provenance

A conference was organized in Zaragoza, Spain, by Pedro Marijuan in 1999 to mark the centenary of the Spanish neuroanatomist Ramón y Cajal's *Textura del Sistema Nervioso del hombre.* Although Cajal wrote very little on consciousness in this great (and still relevant) classic of neuroanatomy, the theme of the conference was *Cajal and Consciousness: Scientific Approaches to Consciousness.* My article from the 1999 meeting was published as (Arbib, 2001), with an abridged version coming out recently (Arbib, 2014) in a special issue of the *Journal of Integrative Neuroscience* on the so-called "hard problem of consciousness." However, it must be noted that I do not claim to solve what some people call the *hard* problem of consciousness but instead focus on the *not at all easy* problem of how, given the presence of consciousness in ancestral species, humans came to have a consciousness with characteristics unshared by other creatures. The present paper combines some of the material from the previous papers with new material that should make it more relevant to the volume of which it is part, although it is not at the biophysical level suggested by the book's title, *Biophysics of Consciousness: A Foundational Approach*.

Acknowledgments

The present chapter is based in part on work supported by the National Science Foundation under Grant No. BCS-1343544 "INSPIRE Track 1: Action, Vision and Language, and their Brain Mechanisms in Evolutionary Relationship" (Michael A. Arbib, Principal Investigator). My thanks to Marsha Kinder and three anonymous referees for their constructive responses to an earlier draft.

References

Adolphs, R. (2009) The social brain: Neural basis of social knowledge. *Annu. Rev. Psychol.,* **60**, 693–716.

Arbib, M. A. (1972) Consciousness: The secondary role of language. *J. Philosophy,* **69**, 579–591.

Arbib, M. A. (1978) Segmentation, schemas, and cooperative computation. In: S. A. Levin, ed. *Studies in Mathematical Biology, Part I.* Providence, RI: Mathematical Association of America, pp. 118–155.

Arbib, M. A. (1985) *In Search of the Person: Philosophical Explorations in Cognitive Science*. Amherst, MA: University of Massachusetts Press.

Arbib, M. A. (1989a) *The Metaphorical Brain 2: Neural Networks and Beyond*. New York: Wiley-Interscience.

Arbib, M. A. (1989b) Modularity, schemas and neurons: A critique of Fodor. In: P. Slezak and W.R. Albury, eds. *Computers, Brains and Minds*. Netherlands: Springer, pp. 193–219.

Arbib, M. A. (1992) Schema theory. In: S. Shapiro, ed. *The Encyclopedia of Artificial Intelligence*. New York: Wiley-Interscience, pp. 1427–1443.

Arbib, M. A. (2001) Co-evolution of human consciousness and language. In: P.C. Marijuan, ed. *Cajal and Consciousness: Scientific Approaches to Consciousness on the Centennial of Ramon y Cajal's Textura*, pp. 195–220.

Arbib, M. A. (2002) The mirror system, imitation, and the evolution of language. In: K. Dautenhahn and C. L. Nehaniv, eds. *Imitation in Animals and Artifacts. Complex Adaptive Systems*. Cambridge, MA: The MIT Press, pp. 229–280.

Arbib, M. A. (2005a) Beware the passionate robot. In: J.-M. Fellous and M. A. Arbib, eds. *Who Needs Emotions? The Brain Meets the Robot*. New York: Oxford University Press.

Arbib, M. A. (2005b) From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics. *The Behavioral and Brain Sciences,* **28**, 105–124.

Arbib, M. A. (2007) Autism — more than the mirror system. *Clin. Neuropsych.,* **4**, 208–222.

Arbib, M. A. (2011) Niche construction and the evolution of language: was territory scavenging the one key factor? *Interact. Stud.,* **12**, 162–193.

Arbib, M. A. (2012) *How the Brain Got Language: The Mirror System Hypothesis*. New York and Oxford: Oxford University Press.

Arbib, M. A. (2013a) Complex imitation and the language-ready brain. *Lang. Cogn.,* **5**.

Arbib, M. A. (2013b) Neurons, schemas, persons and society — revisited. In: G. Auletta, I. Colagè and M. Jeannerod, eds. *Brains Top Down: Is Top-Down Causation Challenging Neuroscience?* Singapore: World Scientific, pp. 57–87.

Arbib, M. A. (2013c) Précis of how the brain got language: the mirror system hypothesis. *Lang. Linguist. Compass,* **5**.

Arbib, M. A., ed. (2013d) *Language, Music and the Brain, a Mysterious Relationship. Strüngmann Forum Reports*, Vol. 10. Cambridge, MA: The MIT Press.

Arbib, M. A. (2014) Co-evolution of human consciousness and language (revisited). *J. Integra. Neurosci.,* **13**, 187–200.

Arbib, M. A. & Fellous, J. M. (2004) Emotions: From brain to robot. *Trends Cogn. Sci.,* **8**, 554–561.

Arbib, M. A. & Hesse, M. B. (1986) *The Construction of Reality*. Cambridge: Cambridge University Press.

Arbib, M. A. & Iriki, A. (2013) Evolving the language- and music-ready brain. In: M.A. Arbib, ed. *Language, Music, and the Brain: A Mysterious Relationship, Strüngmann Forum Reports,* Vol. 10. Cambridge, MA: MIT Press, pp. 359–375.

Arbib, M. A. & Liaw, J.-S. (1995) Sensorimotor transformations in the worlds of frogs and robots. *Artif. Intell.,* **72**, 53–79.

Arbib, M. A. & Rizzolatti, G. (1997) Neural expectations: A possible evolutionary path from manual skills to language. *Commun. Cogn.,* **29**, 393–424.

Barnard, A. (2013) Cognitive and social aspects of language origins. In: C. Lefebvre, B. Comrie and H. Cohen, eds. *New Perspectives on the Origins of Language* Amesterdam/Philadelphia: John Benjamins, pp. 53–71.

Bickerton, D. (1995) *Language and Human Behavior*. Seattle: University of Washington Press.

Bickerton, D. (2009) *Adam's Tongue. How Humans Made Language, How Language Made Humans*. New York: Hill & Wang.

Chomsky, N. (1980) *Rules and Representations*. New York: Columbia University Press.

Dalgleish, T. (2009) The thinking person's emotion theorist: A comment on Bartlett's feeling, imaging, and thinking. *Br. J. Psychol.,* **100**, 199–201.

Damasio, A. (1994) *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Putnam Publishing.

Damasio, A. (1999) *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. San Diego, CA: Harvest/Harcourt.

Damasio, A. (2012) *Self Comes to Mind: Constructing the Conscious Brain*. New York: Vintage.

Dapretto, M., Davies, M.S., Pfeifer, J.H., Scott, A.A., Sigman, M., Bookheimer, S.Y. & Lacoboni, M. (2006) Understanding emotions in others: Mirror neuron dysfunction in children with autism spectrum disorders. *Nat. Neurosci.,* **9**, 28–30.

Darwin, C. (1872) *The Expression of the Emotions in Man and Animals (republished in* 1965). Chicago: University of Chicago Press.

Dehaene, S. & Changeux, J.-P. (2011) Experimental and theoretical approaches to conscious processing. *Neuron,* **70**, 200–227.

di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V. & Rizzolatti, G. (1992) Understanding motor events: A neurophysiological study. *Exp. Brain Res.,* **91**, 176–180.

Dominey, P. F. (2013) The tip of the language iceberg. *Lang. Cogn.,* **5**, 177–187.

Elman, J. L. (2009) On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cogn. Sci.,* **33**, 547–582.

Elman, J. L., Hare, M. & McRae, K. (2004) Cues, constraints, and competition in sentence processing. *Beyond nature-nurture: Essays in honor of Elizabeth Bates*, 111–138.

Fellous, J.-M., & Arbib, M. A. (Eds.). (2005). *Who Needs Emotions: The Brain Meets the Robot*. Oxford, New York: Oxford University Press.

Fellous, J.-M. & Ledoux, J. E. (2005) Toward basic principles for emotional processing: what the fearful brain tells the robot. In: J.-M. Fellous and M. A. Arbib, eds. *Who Needs Emotions: The Brain Meets the Robot*. Oxford, New York: Oxford University Press, pp. 79–115.

Gallese, V. & Goldman, A. (1998) Mirror neurons and the simulation theory of mind-reading. *Trends Cognit. Sci.,* **2**, 493–501.

Gasper, K. & Clore, G. L. (2002) Attending to the big picture: Mood and global versus local processing of visual information. *Psychol. Sci.,* **13**, 34–40.

Gazzola, V., Aziz-Zadeh, L. & Keysers, C. (2006) Empathy and the somatotopic auditory mirror system in humans. *Curr. Biol.*, **16**, 1824–1829.

Goodale, M. A., Milner, A. D., Jakobson, L. S. & Carey, D. P. (1991) A neurological dissociation between perceiving objects and grasping them. *Nature*, **349**, 154–156.

Grafton, S. T., Arbib, M. A., Fadiga, L. & Rizzolatti, G. (1996) Localization of grasp representations in humans by positron emission tomography. 2. Observation compared with imagination. *Exp. Brain Res.*, **112**, 103–111.

Hameroff, S. & Penrose, R. (2014) Consciousness in the universe: A review of the 'Orch OR' theory. *Phys. Life Rev.*, **11**, 39–78, doi: http://dx.doi.org/10.1016/j plrev.2013.08.002.

Head, H. & Holmes, G. (1911) Sensory disturbances from cerebral lesions. *Brain*, **34**, 102–254.

Holmes, G. (1939) The cerebellum of man. *Brain*, **62**, 1–30.

Hubel, D. H. & Wiesel, T. N. (1959) Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.*, **148**, 574–591.

Humphrey, N. K. (1970) What the frog's eye tells the monkey's brain. *Brain Behav. Evol.*, **3**, 324–337.

Iriki, A. & Sakura, O. (2008) The neuroscience of primate intellectual evolution: Natural selection and passive and intentional niche construction. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **363**, 2229–2241.

Iriki, A. & Taoka, M. (2012) Triadic (ecological, neural, cognitive) niche construction: A scenario of human brain evolution extrapolating tool use and language from the control of reaching actions. *Philos. Trans. R. Soc. B: Biol. Sci.*, **367**, 10–23.

Ito, M. (2012) *The Cerebellum: Brain for an Implicit Self*. Upper Saddle River, NJ: FT Press.

Jackson, J. H. (1874) On the nature of the duality of the brain. *Med. Press Circular*, **1,** 19–63.

Jackson, J. H. (1878–1879) On affections of speech from disease of the brain. *Brain*, **1**, 304–330.

Jeannerod, M. (2005) How do we decipher others' minds? In: J.-M. Fellous and M. A. Arbib, eds. *Who Needs Emotions: The Brain Meets the Robot*. Oxford, New York: Oxford University Press, pp. 147–169.

Jeannerod, M., Decety, J. & Michel, F. (1994) Impairment of grasping movements following a bilateral posterior parietal lesion. *Neuropsychologia*, **32**, 369–380.

Kilmer, W. L., McCulloch, W.S. & Blum, J. (1969) A model of the vertebrate central command system. *Int. J. Man-Machine Studies*, **1,** 279–309.

LeDoux, J. (2012) Rethinking the emotional brain. *Neuron*, **73**, 653–676, doi: http://dx.doi.org/10.1016/j.neuron.2012.02.004.

Ledoux, J. E. (2000) Emotion circuits in the brain. *Annu. Rev. Neurosci.*, **23**, 155–184.

Lee, P. (2000) When is 'linguistic relativity' Whorf's linguistic relativity? In: M. Pütz and M. Verspoor, eds. *Explorations in Linguistic Relativity.* John Benjamins, pp. 45–68.

Lettvin, J. Y., Maturana, H., McCulloch, W. S. & Pitts, W. H. (1959) What the frog's eye tells the frog brain. *Proc. IRE,* **47**, 1940–1951.

Levinson, S. C. (2013) Cross-Cultural Universals and Communication Structures. In: M.A. Arbib, ed. *Language, Music, and the Brain: A Mysterious Relationship. Strüngmann Forum Reports, Vol.* 10. Cambridge, MA: MIT Press.

Maravita, A. & Iriki, A. (2004) Tools for the body (schema). *Trends Cogn. Sci.,* **8**, 79–86.

Merker, B. (2007a) Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behav. Brain Sci.* **30**, 63–81.

Merker, B. (2007b) Grounding consciousness: The mesodiencephalon as thalamocortical base. *Behav. Brain Sci.* **30**, 110–120.

O'Regan, J. K. (2011) *Why Red doesn't Sound like a Bell: Understanding the Feel of Consciousness*. Oxford: Oxford University Press.

O'Regan, J. K. (2012) How to build a robot that is conscious and feels. *Mind. Mach.,* **22**, 117–136, doi: 10.1007/s11023-012-9279-x.

O'Regan, J. K. & Block, N. (2012) Discussion of J. Kevin O'Regan's "why red doesn't sound like a bell: Understanding the feel of consciousness." *Rev. Philos. Psychol.,* **3**, 89–108, doi: 10.1007/s13164-012-0090-7.

Odling-Smee, F. J., Laland, K. N. & Feldman, M. W. (2003) *Niche Construction: The Neglected Process in Evolution*. Princeton, NJ: Princeton University Press.

Olds, J. (1969) The central nervous system and the reinforcement of behavior. *Am. Psychol.,* **24**, 114–132.

Ortony, A., Clore, G. L. & Collins, A. (1988) *The Cognitive Structure of Emotions*. New York: Cambridge University Press.

Ortony, A., Norman, D. & Revelle, W. (2005) Affect and proto-affect in effective functioning. In: J.-M. Fellous and M. A. Arbib, eds. *Who needs emotions* New York and Oxford: Oxford University Press, pp. 173–202.

Panksepp, J. (1998) The periconscious substrates of consciousness: Affective states and the evolutionary origins of the self. *J. Consciousness Stud.,* **5**, 566–582.

Pinker, S. (2010) The cognitive niche: Coevolution of intelligence, sociality, and language. *Proc. Natl. Acad. Sci.,* **107**(Suppl2), 8993–8999, doi: 10.1073/pnas.0914630107.

Pribram, K. H. (1960) A review of theory in physiological psychology. *Annu. Rev. Psychol.,* **11**, 1–40.

Ramachandran, V. S. (2011) *The Tell-Tale Brain: A Neuroscientist's Quest for What Makes Us Human*. London: W. W. Norton & Company.

Ramachandran, V. S. & Oberman, L. M. (2006) Broken mirrors: A theory of autism. *Sci. Am.,* **295**, 62–69.

Rizzolatti, G. & Arbib, M. A. (1998) Language within our grasp. *Trends Neurosci.,* **21**, 188–194.

Rizzolatti, G., Fadiga, L., Gallese, V. & Fogassi, L. (1996a) Premotor cortex and the recognition of motor actions. *Cogn. Brain Res.,* **3**, 131–141.

Rizzolatti, G., Fadiga, L., Matelli, M., Bettinardi, V., Paulesu, E., Perani, D. & Fazio, F. (1996b) Localization of grasp representations in humans by PET: 1. Observation versus execution. *Exp. Brain Res.,* **111**, 246–252.

Rizzolatti, G., Fogassi, L. & Gallese, V. (2001) Neurophysiological mechanisms underlying the understanding and imitation of action. *Nat. Rev. Neurosci.*, **2**, 661–670.

Rolls, E. T. (2000) Neurophysiology and functions of the primate amygdala, and the neural basis of emotion. In: J.P. Aggleton, ed. *The Amygdala: A Functional Analysis*. 2nd edn. Oxford: Oxford University Press, pp. 447–478.

Rolls, E. T. (2005) What are emotions, why do we have emotions, and what is their computational role in the brain? In: J.-M. Fellous and M.A. Arbib, eds. *Who Needs Emotions: The Brain Meets the Robot*. Oxford, New York: Oxford University Press, pp. 117–146.

Sarraute, N. (1984) *Childhood (Translated by Barbara Wright in consultation with the author)*. George Braziller.

Schultz, W. (2000) Multiple reward signals in the brain. *Nat. Rev. Neurosci.*, **1**, 199–207.

Seifert, U., Verschure, P. F. M. J., Arbib, M. A., Cohen, A. J., Fogassi, L., Fritz, T. & Scherer, K. (2013) Semantics of internal and external worlds. In: M.A. Arbib, ed. *Language, Music, and the Brain: A Mysterious Relationship, Strüngmann Forum Reports*, Vol. 10. Cambridge, MA: MIT Press.

Sloman, A., Chrisley, R. & Scheutz, M. (2005) The architectural basis of affective states and processes. In: J. M. Fellous and M. A. Arbib, eds. *Who Needs Emotions*. New York & Oxford: Oxford University Press, pp. 203–244.

Stoerig, P. & Cowey, A. (1997) Blindsight in man and monkey. *Brain*, **120**, 535–559, doi: 10.1093/brain/120.3.535.

Suddendorf, T. & Corballis, M. C. (2007) The evolution of foresight: What is mental time travel, and is it unique to humans? *Behav. Brain Sci.*, **30**, 299–351.

Tomasello, M., Carpenter, M., Call, J., Behne, T. & Moll, H. (2005) Understanding and sharing intentions: The origins of cultural cognition. *Behav. Brain Sci.*, **28**, 675–691.

Weiskrantz, L. (1974) The interaction between occipital and temporal cortex in vision: An overview. In: F.O. Schmitt and F.G. Worden, eds. *The Neurosciences Third Study Program*. Cambridge, MA: The MIT Press, pp. 189–204.

Whitehead, C. (2010) The culture ready brain. *Soc. Cogn. Affect. Neurosci.*, **5**, 168–179, doi: 10.1093/scan/nsq036.

Whorf, B. L. (1937) Discussion of Hopi linguistics. In: J.B. Carroll, ed. *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. Cambridge, MA: MIT Press, pp. 102–111.

Wicker, B., Keysers, C., Plailly, J., Royet, J. P., Gallese, V. & Rizzolatti, G. (2003) Both of us disgusted in My insula: The common neural basis of seeing and feeling disgust. *Neuron*, **40**, 655–664.

Williams, J. H. G. (2008) Self-other relations in social development and autism: Multiple roles for mirror neurons and other brain bases. *Autism Res.*, **1**, 73–90.

Williams, J. H. G., Whiten, A., Suddendorf, T. & Perrett, D. I. (2001) Imitation, mirror neurons and autism. *Neurosci. Biobehav. Rev.*, **25**, 287–295.