Artificial Intelligence 60 (1993) 313–356
Elsevier

ARTINT 1056

Book Review

The biology of consciousness: comparative review of Israel Rosenfield, *The Strange, Familiar, and Forgotten: An Anatomy of Consciousness*[*] and Gerald M. Edelman, *Bright Air, Brilliant Fire: On the Matter of the Mind*[**]

William J. Clancey

Institute for Research on Learning, 2550 Hanover Street, Palo Alto, CA 94304, USA

Revised January 1993

1. Relating cognition to biology

For many years, most AI researchers and cognitive scientists have reserved the topic of consciousness for after dinner conversation. Like "intuition", the idea of consciousness appeared to be too vague or general to be a good starting place for understanding cognition. Work on narrowly-defined problems in specialized domains such as medicine and manufacturing focused

[*] (Alfred A. Knopf, New York,1992); xi + 157 pages.

[**] (Basic Books, New York, 1992); xvi + 280 pages.

Correspondence to: W.J. Clancey, Institute for Research on Learning, 2550 Hanover Street, Palo Alto, CA 94304, USA. Telephone: (415) 496 7925. Fax: (415) 496 7957. E-mail: bill_clancey@irl.com.

our concerns on the nature of representation, memory, strategies for problem solving, and learning. Some writers, notably Ornstein [34] and Hofstadter [27], continued to explore the ideas, but implications for cognitive modeling were unclear, suggesting neither experiments, nor new computational mechanisms.

But the time has arrived for raising consciousness in cognitive science. Books by Edelman, Rosenfield, Dennett, Varela, and others have appeared almost simultaneously, with a strikingly common theme: Biological and psychological evidence suggests that better understanding of consciousness is not only possible, but necessary if we are to improve our understanding of cognition. This evidence varies considerably, ranging from how neurological structures develop, the effects of neural dysfunctions on human behavior, perceptual illusions, the evolution of the human species, and the philosophy of language. In this comparative review, I consider the work of Edelman and Rosenfield. Taken together, these books may stimulate a broader view of intelligence, give further credence to the situated cognition view of language, and provide a more biological basis for "neural net" approaches. Prior work by Putnam, Dreyfus, and Winograd, to name a few previous critics, may also appear less threatening or less nonsensical when argued in neurophysiological terms.

2. Overview of *The Strange, Familiar, and Forgotten*

Rosenfield's psychological analysis of human experience motivates Edelman's more detailed neurological models of the brain, so I will consider *The Strange, Familiar, and Forgotten* first. This book is a significant contribution to cognitive science literature, in the interpretive, historical style of Sacks [43] and Luria [31]. Like Sacks and Luria, Rosenfield is an MD with a historian's bent. With a PhD in "intellectual history", Rosenfield attempts to make sense of clinical neuropsychology research (which he often translates from French sources). Reinterpreting past work, he applies a developmental perspective that learning occurs with every human interaction.

Like Sacks and Luria, Rosenfield uses historically well-documented cases to illustrate and contrast theories of memory, learning, and consciousness. Like them, he provides an ethnographic perspective on the patient, not merely as a patient with a lesion but as *a person struggling to make sense of emotional, physical, and social experience.* He considers not only laboratory evidence of the patients' verbal and perceptual behavior, but the stories they tell about their social life and mental experience.

Building on simple observations and comparison across cases, Rosenfield provides a broad view of experiences that theories of memory especially must address. For example, Mr. March could move his left hand, but only when told to do so (p. 58), and otherwise seemed not to relate to it as his own. When a nurse made her hand appear to be his left hand, he casually made sense of the situation:

> Asked where his own ring was, he said, "It's been taken away from me."
> And why was he now wearing a bracelet? "It has been put on me."
> "But this hand is all white and not as hairy as your own."
> "It's like that because it is paralyzed." (p. 59)

Rosenfield summarizes his thesis that dysfunctions are usually exhibited in the context of ongoing sense-making behavior:

> The neurologists' attempts to derive brain function from clinical reports of brain-damaged patients have too often overlooked the fact that the verbal reports of these patients are *conscious* reports ... limited awareness causes not the "loss" of words but an inability to make a certain sense of them and thus to use them in conventional ways ... in a concentration on the idea that individual functions had been lost or damaged in brain lesions, important and subtle symptoms went unexplained. Yet these symptoms were part of the patients' *conscious* states, and they suggest a broader functional breakdown than the view of compartmentalized functions allows. (p. 140)

Rosenfield argues that we need to better understand the nature of consciousness as a process, an activity in its own right, not as a side-effect, but an ongoing accomplishment. Related work by Greenwald [26] argues for the need to "explain one particularly intriguing 'emergent' property of the self system—its tendency (in the normal case) to perceive itself as unitary and real". We need a global view of the dynamics of sense-making, of creating an integrating view—in our conscious behavior—of our self-image. This parallels the emphasis in situated cognition (e.g., Suchman's [50] study of the use of plans) on *representing* as occurring in sequences of interactive behavior over time, within some ongoing situation, as opposed to being disembodied manipulation of calculi, in some timeless hidden (and subconscious) place inside the brain. Rosenfield and Edelman help us understand *why all reasoning isn't subconscious*: Sense-making (telling causally-connected stories) involves relating what you are currently doing to what you have experienced in the past and what you expect will occur in the future. Significantly, "relating" occurs without necessarily representing such relations in words, but rather directly, via neural feedback loops that couple perception, past (non-linguistic) conceptualizations, and bodily movements.

What is a patient *doing* when he says that last Saturday he was in the city of La Rochelle, when he has no idea what day was yesterday and denies that he is still in La Rochelle today? Rosenfield closely examines such story-telling behaviors, revealing that patients are not merely retrieving facts from memory, but revealing how they make sense of experience. By examining what can go wrong—in maintaining a sense of continuous time, an integrated personality and body image, and abstract categorical relations—we can understand better how consciousness structures everyday experience.

Rosenfield's view of the brain is consistent with cognitive science descriptions of behavior patterns (scripts, grammars), in so far as he acknowledges that such patterns are real psychological phenomena that need to be explained. Yet, he insists on an alternative view of neurological mechanism, by which observed behavior patterns are the product of interactions at both social and neural levels. This idea of *dialectic organization* is important in biology and anthropology, but quite different from the mechanisms designed by most engineers, computer scientists, and cognitive modelers. Stephen Jay Gould provides a useful introduction:

> Thus, we cannot factor a complex social situation into so much biology on one side, and so much culture on the other. We must seek to understand the emergent and irreducible properties arising from an inextricable interpenetration of genes and environments. In short, we must use what so many great thinkers call, but American fashion dismisses as political rhetoric from the other side, a dialectical approach.
>
> ... the three classical laws of dialectics embody a holistic vision that views change as interaction among components of complete systems, and sees the components themselves not as a priori entities, but as both the products of and the inputs to the system. Thus the law of "interpenetrating opposites" records the inextricable interdependence of components; the "transformation of quantity to quality" defends a systems-based view of change that translates incremental inputs into alterations of state; and the "negation of negation" describes the direction given to history because complex systems cannot revert exactly to previous states. [25, pp. 153–154]

According to this dialectical perspective, neural processes activate and are generalized within larger neural and social coordinations which they constitute, yet which create them. Areas of the brain are specialized, but the degree of modularity and stability is different from the labeled memory structures and independently invokable subroutines of most cognitive models. Areas of the brain aren't merely accessed or activated, but *organize each* *other within complete circuits* (as emphasized in Dewey's 1896 criticism of stimulus–response theories [17]). According to Edelman's model of the brain, these circuits are themselves generalizations involving bi-directional recategorizations at perceptual, sequential, conceptual, and linguistic levels.

Like Edelman, Rosenfield emphasizes the "interpenetrating" multiple levels of individual development, species evolution, and the interaction of cultural and neural processes. But in the more narrow style of cognitive neuropsychology, he focuses on what abnormal behavior reveals about normal function. In order to explain dysfunctions, as well as the openness and subjectivity of categories in everyday life, Rosenfield argues for a brain that continuously and dynamically reorganizes how it responds to stimuli (p. 134). Rather than retrieving and matching discrete structures or procedures, the brain composes itself in-line, in the very process of coordinating sensation and motion (hence behavior is "situated"). By Dewey's analysis, perceptual and motor processes in the brain are configuring each other without intervening subconscious "deliberation" [13].

Most interpretations of patients with dysfunctions (e.g., an inability to speak certain kinds of words) have postulated isolated memory or knowledge centers for different kinds of subsystems: auditory, visual, motor. The "diagram makers", exemplified by Charcot in the mid-1800s, drew pictures of the brain with "centers", linking memory and parts of the body. Dysfunctions were explained as loss of memory, that is, loss of specific knowledge stored in the brain. Rosenfield claims instead that memory loss in a brain-damaged patient is not the loss of a "memory trace", but evidence of a restructuring of how the brain operates. That is, we are not observing a primary, isolated "loss", but a secondary *process of reorganization* for the sake of sustaining self-image:

> Memory loss in the brain-damaged patient is ... evidence of a restructuring of the patient's conscious knowledge, a restructuring of the patient's relation to his or her surroundings. The brain has mechanisms for establishing this relation—that is the ultimate significance of the pathological findings—and the most important significance of these mechanisms is consciousness. With brain damage, function is altered, certain brain processes are no longer possible, and consequently consciousness, too, is altered. (p. 22)

> Patients with brain damage are confused when they fail to recognize and remember, and it is this confused, altered awareness, as much as any specific failures of memory, that is symptomatic of their illness. (p. 34)

The kinds of *confusion* reveal the normal function of the brain in coordinating present awareness with previous experience. Being conscious is being engaged in the act of sense-making; brain-damaged patients exhibit "a breakdown in the mechanisms of consciousness. A patient's state of confusion is no more to be ignored than his failure to recognize, say, his home. Memory, recognition, and consciousness are all part of the same process." (p. 35) Self-reference is integral to sense-making, and it is grounded in bodily experience. In people, there can be no "normal" understanding of language, no sense of time, no personality without a sustained, coherent self-image. The cases of brain-damaged patients support this view, but ultimately Edelman's architectural arguments provide the needed implementation-level support.

The first section of the book recapitulates some of the analysis from Rosenfield's *Invention of Memory* [11,41], but elaborated from the perspective of conscious experience:

> Our perceptions are part of a "stream of consciousness," part of a continuity of experience that the neuroscientific models and descriptions fail to capture; their categories of color, say, or smell, or sound, or motion are discrete entities independent of time. ... a sense of consciousness comes precisely from the *flow* of perceptions, from the relations among them (both spatial and temporal), from the dynamic but constant relation to them as governed by one unique personal perspective sustained throughout a conscious life.... Compared to it, units of "knowledge" such as we can transmit or record in books or images are but instant snapshots taken in a dynamic flow of uncontainable, unrepeatable, and inexpressible experience. And it is an unwarranted mistake to associate these snapshots with material "stored" in the brain. (p. 6)

By suggesting that memory was a place or a capability physically separated and distinct from the function of the brain (in speaking, moving, reasoning), Wernicke and others may have "falsified our understanding of numerous clinical disorders and of brain function in general" (p. 22). Dennett [15] provides a similar analysis:

> The consensus of cognitive science ... is that *over there* we have the long-term memory ... and *over here* we have the workspace or working memory, where the thinking happens .... And yet there are no two places in the brain to house these two facilities. The only place in the brain that is a plausible home for either of these separate functions is the whole cortex—not two places side by side but one large place. (pp. 270–271)

Rosenfield's model, in which memory is integral to skills, resembles theories such as Schank's dynamic memory (MOPS) and case-based reasoning, in which experiences and generalizations are integrated. However, these models postulate that declarative facts are stored as discrete units, even if they are linked to how they have been "accessed" or "used" in the past. Rosenfield and Edelman consistently emphasize that the brain operates only procedurally, with no storage of semantic information, either declarative or programmatic, as discrete linguistic structures [11].

Unlike Edelman, Rosenfield makes no attempt to address the AI audience directly. His statements require some reformulation to bring home the insights. For example, Rosenfield says, "No machine is *troubled* by, or even *intrigued* by, feelings of certainty that appear contradictory" (p. 12). Yet, AI researchers cite examples of how a program detects contradictory conclusions and uses that information. Without further discussion, it is unclear how being troubled is an essential part of creating new goals and values.

Providing a convincing case requires understanding the reader's point of view well enough to anticipate rebuttals. To this end, I re-present Rosenfield's key cases and contrast his analysis with other cognitive science explanations. The central themes of his analysis are: non-localization of function, the nature and role of self-reference (subjectivity), the origin and sense of time in remembering, the relational nature of linguistic categories, and the problem of multiple personalities. I reorder these topics in order to more clearly convey the neural-architectural implications.

2.1. Multiple personalities

The process of *sustaining a self-image* is illustrated by people with multiple personalities. Rosenfield argues such experiences are caused by pathology that *limits* neural organizational processes.

For example, "Mary Reynolds could, at different times, call the same animal a 'black hog' or a 'bear'." This behavior was integrated with alternative personalities, one daring and cheerful, the other fearsome and melancholic. From the standpoint of cognitive modeling based on stored linguistic schemas, there would be two memories of facts and skills, two coherent subconscious sets of representational structures and procedures. Rosenfield argues instead:

> There cannot be unconscious "traces" of these conscious states, since they require a dynamic organization that, given the complexity of the processes (the immediate, the past, and self-reference), are not reproducible. But what *are* more or less reproducible are the ways in which the brain organizes itself; certain pathologies limit the organizational processes, not the accessibility or inaccessibility of memories. (p. 129)

Again, neural structures coordinating what we say, imagine, feel, and how we move are activated "in place", as we are in the process of speaking, feeling, moving. Saying that some memories are forgotten and others recalled suggests a process of search and matching for relevancy; instead, the brain directly reorganizes itself *on a global basis*, not merely filtering or "reinterpreting" sensations, but physically recoordinating how perception and conceptualization occur.

Referring to another multiple personality, Rosenfield says,

> So there is no "Ansel" organized as such in "Arthur's" brain, or vice versa. Rather the single brain organizes itself as if it were Ansel (and there had never been an Arthur), then vice versa, because under certain conditions this damaged brain is reorganizing its way of responding to stimuli, the nature of its relations to the world.... Knowledge is the brain's ability to organize itself in particular ways at particular times. (p. 123)

We sometimes experience such figure–ground switches in our own experience: "Yesterday's 'friend' is today's 'objectionable person'." In normal people recategorization is gradual. In the patient with multiple personalities, "None of his or her personalities fully 'fits' the dynamic experience of everyday life: one personality will recognize family and friends; another will treat them as strangers and enemies" (p. 123). There are "too few 'selves'" (p. 138), a confined repertoire of fixations. Each personality appears one-dimensional; the reorganizations are disjoint. There is no gradation of "somewhat remembering" or "somewhat being able to control" the other personalities. This clinical evidence supports the neurobiological model of modular co-determination provided by Edelman, which we will discuss later.

2.2. *Self-reference*

Rosenfield's analysis of Mr. March's left-side discoordination supports Lakoff's [29] view that understanding of the world is grounded in and emerges from the dynamics of body movements. Body movements serve as a frame of reference by which stimuli are organized, and upon which more complex coordinations and categorizations are built. For the infant, consciousness only develops after "genetically determined reflexes" (p. 61) provide initial experience upon which stimuli can be systematically experienced and organized. The relation between new and old (what Edelman calls "the remembered present") is experienced as a sense of self-familiarity during activity itself, forming the basis of self-awareness and ultimately personality.

The nature of self-reference is revealed by patients with loss of awareness of their limbs ("alien limbs"), as well as the changed experience of people who become blind. For example, Hull suggests that after becoming blind he has difficulty in recollecting prior experiences that involve seeing. Rosenfield's interpretation is that neurological structures involved in seeing are now difficult to coordinate with Hull's present experience. His memories are not lost, but his visual self-reference is limited. That is, remembering is a kind of coordination process. Hull's present self-image is *visually restricted*: "There is no extension of awareness into space ... I am dissolving. I am no longer concentrated in a particular location ..." (quoted by Rosenfield, p. 64). Without an *ongoing* visual frame of reference, he is unable to establish relations to his *prior* visual experience.

Again, recollecting isn't retrieving and reciting the contents of memory, but a dynamic process of establishing "relations to one's present self" (p. 66). "Establishing relations" means physically integrating previous neural activations with currently active neural processes. A stored linguistic schema model fits Hull's experience less well than a model of memory based on physical processes of bodily coordination.

The case of Oliver Sacks' [43] alien leg (paralyzed and without feeling because of an accident) also reveals self-referential aspects of awareness. Upon seeing his leg in his bed and not recognizing it, Sacks actually tossed it out of the bed, landing on the floor. Rosenfield emphasizes that "seeing is not by itself 'knowing' and that the lack of inner self-reference, together with the incontrovertible sight of the leg, therefore created a paradoxical relation to it" (p. 53). Sacks is not just *sensing* his leg, for he can clearly see it and recognizes that it is a leg. Instead, Sacks is relating his categorizations to his ongoing sense of himself and his surroundings. If we see a strange object in our bed (especially an unfamiliar leg), we move to throw it out. Rosenfield argues that *we cannot separate categorizing from this ongoing process of sustaining the self versus non-self relation.* Both Rosenfield and Edelman argue that most cognitive models and AI programs lack self-reference, or view it as a secondary reflection after behavior occurs. Edelman's models suggest that self-reference involves neurological feedback between levels of categorization, including feedforward relations between higher-level coordinating and lower-level perceptual categorizations.

Another patient studied by Charcot, Monsieur A, lost his ability to recognize shapes and colors. He could no longer draw or visualize images (both of which he previously did extraordinarily well) or even recognize his family. As often occurs in these cases, Monsieur A was now, in his words, "less susceptible to sorrow or psychological pain". This *diminished sense of pleasure and pain indicates a change in self-reference*, of awareness of the self. Oddly, Monsieur A could speak and answer questions and continued his work and everyday life in a somewhat disjointed way. But he was unable to establish a relation between words and his sensory experience. He understood words only in their abstract relations. As in Hull's case, this inability to perceive now impaired his ability to remember; as Monsieur A put it, "Today I can remember things only if I say them to myself, while in the past I had only to photograph them with my sight" (p. 93). This suggests again that remembering is integrated with sensory experience, that *remembering is a form of perceiving.*

Early work by the "diagram makers" viewed neural lesions in terms of cutting off areas of the brain, such that stored images, word definitions, or the like are inaccessible. Contradicting Charcot, Rosenfield claims that Monsieur A had not lost specific visual memories, but his ability to integrate *present* visual experience—to establish a present sense of himself—that included immediate and practical relation with colors and shapes. Semantic content doesn't reside in a store of linguistic categorizations, but in the relation of categorizations to each other. Indeed, *every categorization is a dynamic relation between neural processes.* Edelman's model suggests that in Monsieur A neural maps that ordinarily relate different subsystems in the brain are unable to actively coordinate his visual sensory stimuli with ongoing conceptualization of experience. Experiments show that sensory categorization may still be occurring (e.g., some patients unable to recognize friends and family may exhibit galvanic skin responses) (p. 123). But conscious awareness of sensation requires establishing a *relation* with the current conceptualization of the self.

The process of sustaining self-coherence has a holistic aspect, such that loss of any one sensory modality has global effects on memory and personality. Again, this argues for consciousness not as a side-effect of a discrete assembly of components, but as the business of the brain as it coordinates past activation relations with ongoing perception and movement.

2.3. Time

A sense of time is inherently relational. Time is another manifestation of self-reference, awareness that present experience bears a relation to what we experienced before. Rosenfield argues that such feedback is inherently part of ongoing conscious awareness. When it is impaired, not only is memory impaired, but also our ability to learn, to coordinate complex concepts, and to sustain a coherent personality. Again, dysfunction reveals processes that we take for granted in everyday experience and inadequately credit in our theories of cognition.

Mabille and Pitres' patient in 1913, Mr. Baud, provides a good example. When asked if he knows the town of La Rochelle, he replies that he went there some time ago to find a pretty woman. He remembers where he stayed, and says that he never went back. Yet he has been in a hospital in La Rochelle for thirteen years. He also says he has a mistress whom he sees every Saturday. Asked when he last saw her, he responds, "last Saturday".

But again he hasn't left the hospital in all this time (p. 80). Contradicting Mabille and Pitres, Rosenfield tells us that we have no idea what the patient meant by "last Saturday". It could not possibly be a specific Saturday since he has no idea what day today is; Mr. Baud has no specific memory beyond the past twenty seconds.

Oddly, Mr. Baud is a bit like a stored linguistic schema program. He knows how to use words in a conversation, but he has no ongoing, connected experience. He is like a program that has only been living for twenty seconds, but has a stored repertoire of definitions and scripts. He knows the patterns of what he typically does and answers questions logically on this basis: Since he goes every Saturday, he must have gone last Saturday. But from the observer's perspective, which transcends Mr. Baud's twenty-second life span, he lacks a sense of time. Interpreting his case is tricky, because Mr. Baud can't be recalling La Rochelle as we know it if he doesn't acknowledge that he's currently in that town. How in fact, could he be recalling any *place* at all or any *time* at all in the sense that we make sense of our location and temporal experience?

Rosenfield argues that our normal "relation to the world is not sometimes abstract and sometimes immediate, but rather *always both*" (p. 80). To say that Mr. Baud has abstract, long-term memory, but lacks immediate, short-term memory ignores how our attention shifts in normal experience as we relate recollections to what we are experiencing now. "Distant experiences become specific—refer to a specific event in our past—when we can relate them to our present world" (pp. 75–76). Without this ability to coordinate his reminiscences with his present experience, Mr. Baud exhibits a breakdown in an aspect of sense-making, not merely a loss of memory or inability to store recent experiences. His recollections are "odd abstractions, devoid of temporal meaning" (p. 77). That is, a sense of time involves a kind of self-reference that Mr. Baud cannot experience.

Strikingly, this view of meaning goes beyond the idea of "indexicality", previously emphasized in situated cognition (e.g., [1]). Understanding is not just establishing the relation of words like "last Saturday" to the present situation. Knowing the present situation involves having a dynamic sense of self. Without being able to relate my experiences (either past or present) back to *me* (p. 87), my awareness of the past, of history, of memory, and the present situation will be impaired. Put another way, understanding, as well as remembering and reasoning, involve orienting my *self.* If I am confused about who I am, I can't understand what "here" and "now" mean.

Consciousness as a mechanism sustains a relation between our recollections and our ongoing sense of self. Naming, history-telling, and theorizing—integral aspects of sense-making [45]—are ways of establishing relationships in our experience (p. 98). Mr. Baud, lacking an "immediate" relation to his surroundings, can't have an abstract relation either. His recollections are *timeless* in lacking a relation to the present (p. 80). The view that there are isolated functions, such as short and long-term memory, and that one is simply missing, is inadequate. It is Mr. Baud's ability to coordinate neural processes, not access stored facts, that is impaired. This illustrates the thematic contention of Rosenfield and Edelman that cognitive science benefits from a biological re-examination of the nature of memory. In effect, prevalent functional models and computational engines assume a separability of activation and processing that the brain does not employ.

Rosenfield cites studies that suggest that dysfunctions like Mr. Baud's appear to be caused by damage to the hippocampus and associated structures in the limbic system (p. 85), which is "essential for establishing the correlations between the body image and external stimuli that are the basis of consciousness" (p. 86). Edelman's analysis goes further to relate such "primary consciousness" to conceptual categorization by the cortex.

2.4. *Language*

Rosenfield's discussion of language provides a good introduction to Edelman's model. Both authors claim that the nature of the brain's development, coupled with the evolution of language, suggest that grammars are neither innate nor stored as discrete structures in the brain. Speaking a language involves continual recategorization of both the sounds and meanings of words. A stored-rule view of static information is not sufficient to explain how we dynamically understand "different speakers [who] pronounce words differently, and a given speaker may pronounce the same word in a number of different ways" (p. 37). Rosenfield and Edelman argue that we are restructuring previous neural activations directly, not reasoning about features of sounds or using an intermediate, descriptive representation. Crucially, *the same in-place adjustments occur as we conceptualize and understand meanings.*

As an example, Rosenfield describes brain-damaged patients unable to use terms like "red" abstractly, but who can nevertheless perceive and sort objects by color. Again, Rosenfield argues against a "disconnection model" in which concepts like "red" are assumed to be innate categories. This model is still current in cognitive neuropsychology, with claims that different kinds of words such as proper nouns, verbs, prepositions, etc. are stored in different parts of the brain that are genetically functionally specialized. An alternative explanation is that the mechanism by which *relations* are constructed and differences generalized is impaired: The patient "finds puzzling Gelb and Goldstein's insistence that all the variant shades are 'red'" (p. 103). The patient has difficulty forming kinds of *conceptual* categorizations (i.e., coordinating perceptual categorizations), not retrieving facts about colors.

Using examples from infant learning, Rosenfield argues, "Learning a language might well be described as the acquisition of the skill of generalization or categorization" (p. 105). That is, *naming is a sense-making activity*, and processes of categorization build on each other. So, for example, "children first learn the words for size and only later the words for colors" (p. 105).

Crucially, categories are *relations*, as Rosenfield says repeatedly. Edelman calls coherent responses to stimuli "perceptual categorization". But according to Rosenfield, this fails to emphasize that each categorization is a relation to other coherent coordinations (ongoing and previously activated) (p. 83). That is, the meaning of a concept is embodied in the functional relation of ongoing neural processes, themselves constructed from prior coordinations. Bartlett made this same point in 1932:

> It is with remembering as it is with the stroke in a skilled game. We may fancy that we are repeating a series of movements learned a long time before from a text-book or from a teacher. But motion study shows that in fact we build up the stroke afresh on a basis of the immediately preceding balance of postures and the momentary needs of the game. Every time we make it, it has its own characteristics.
>
> [T]here is no reason in the world for regarding these [traces/schemata] as made complete at one moment, stored up somewhere, and then re-excited at some much later moment. [5, p. 211]

Rosenfield provides a useful analogy:

> How categorization of a stimulus is achieved might be best understood by an analogy. Imagine, for example, a group of musicians, let us say a string quartet. As each member of the quartet plays his individual instrument, he both sends to and receives from his fellow musicians "signals" about the sound, volume, rhythm, accent, and tone quality of the music. Each player is carrying on an individual dialogue with the other players, together creating a sound at any given moment. There is no conductor, no central command. So, too, in the brain, local interactions among the brain's maps, their "speaking" back and forth to each other by an exchange of signals, creates a coherent response to a stimulus. The response to the stimulus is not predetermined; local interactions among different parts of the brain give the response its coherence. Just as the shape and overall sound of the quartet's performance is created by the various sounds from moment to moment, so, too, categorizations emerge from the brain's relating one coherent response to another and another. (p. 83)

Instead of a score, the brain's "players" are reenacting their previous roles, improvising their relations to each other in a new composition. Chaos or "oscillation" models of the brain [23] suggest that both local interactions and global effects may be accommodated simultaneously, in real time. Edelman's "maps of maps" provide another top-down organizing mechanism. The key ideas involved in dialectical control and organization are: (1) composition in place (as opposed to use of buffers, copying, or a central place where conceptions are assembled; in terms of the analogy, the musical effect arises and exists only in actual playing), and (2) no intermediate linguistic descriptions in the forms of grammars, scripts, or word definitions (except in so far as the person interactively engages in such representing behavior in cycles of perceiving and acting over time; taking our analogy literally, musical scores are only interpreted in playing over time, not stored and executed internally). Finally, for the ensemble of musicians, as well as the brain, coherence arises because of the *relation* between local and global constraints.

Possible relations are constrained by the available mechanisms in the brain, the evolution of human language as a social process, and the development of the individual. Rosenfield illustrates the generalization process of creating new languages with examples of sign language and Creole. Note that the issue is *how a new language develops*, not how people learn an existing language. In this case, linguists observe a two-generation process by which the first generation of children and adults develops gestures or pidgin, with only a simple grammar if any at all. The younger children of the second generation "abstract (categorize) the gestures of older students, creating from them symbols and more abstract categories of relations among these symbols—a true grammar. An older child may point (gesture) to a rabbit to indicate his subject; a younger one will categorize the pointing gesture as 'rabbit,' and the gesture becomes a symbol" (pp. 110–111). The need for a second generation suggests that neural mechanisms alone are not sufficient, in the individual, for developing a new grammatical language.

In effect, the experience of many different gestures present in the environment becomes categorized into a repeated experience of "gesturing", with an associated typology and ordering of gestures as symbols. A similar analysis is demonstrated by Bamberger's [4] studies of children learning to perceive and use musical tones, not just as integrated parts of a melody, but as named and ordered objects that can be manipulated to produce meaningful sequences. This developmental process illustrates the "brain's constant reworking of its own generalizations" (p. 111).

The patterns between sign language and Creole language development suggest the importance of social sharing of language, as well as neurological constraints that limit an older child's ability to abstract a language beyond its immediate and practical relations. We are reminded of brain-damaged patients who lack a sense of abstract time, color, shape, etc., but can handle particulars in the "here and now". (Again, without a possible concomitant abstract meaning, the particulars have a different sense than experienced by people with abstraction capabilities.)

Rosenfield and Edelman both argue against the existence of specific innate categories or grammar rules, and emphasize the overwhelming importance of cultural influences on what can be accomplished by individuals. However, they agree that the mechanism that enables *grammatical* language to develop involves neurological structures that evolved in the human species and are not found in other animals. Nevertheless, these are "new areas of the brain ... not new principles of mental function" (p. 119). This is supported by Edelman's model, which shows new relations between existing processes of categorization, not a new kind of compositional activation process. Similar arguments are made by Head and Bartlett; more recently, Calvin [9] claims that sequencing control processes for physical movements such as throwing are involved in speech and complex conceptualization. That is, the same kinds of neurological processes may occur in different areas of the brain, becoming specialized for different functions through use. Rosenfield calls this the "holistic" view, in which parts are "not independently specialized, but interdependent" (p. 24). Categorizing areas of the brain establish dynamic, time-sensitive relations to other areas, as opposed to storing discrete representations of words, sounds, meanings, etc. in isolation. This is also what Rosenfield means when he says that functions are not predetermined, either inborn in the infant or as pre-stored responses in the adult.

The primary repertoire of neural interactions, within which categorization and coordination occurs, is not determined by genes, but develops in adolescence through a complex process involving topological constraints, redundant connections, and experiential strengthening. Even the brains of identical twins are wired differently (p. 82). This is of course strong evidence against the idea that specific linguistic rules or categories could be inherent in the brain. Rather the existence of commonalities in human language, known as universal grammar, is evidence of common *transformational* principles by which categories are formed.

Specifically, language adds a new kind of self-reference (p. 119), in which we become explicitly conscious of ourselves, by naming of phenomenological experience, historical accounts, and causal rationalizations [45]. This self-reference required the evolution of a special memory system that "categorized the vocal cord's gestural patterns":

> The brain, *linking these gestures to its nonlinguistic categorizations of its own activities*, and categorizing these linked signals in another special memory system, created the basis for a gestural system that can refer to objects and actions. A developed gestural

> language became a stimulus [internally] and was recategorized into symbols and a true syntax. After sufficient lexical experience, the language was in turn treated as a stimulus by the categorical centers and recategorized; thus language became an independent means of thought, creating the notion of time past, present, and future. (pp. 112–113; emphasis added)

As we will now see, Edelman's models of the brain specify what areas of the brain are involved and their relations to each other.

3. Overview of *Bright Air, Brilliant Fire*

AI researchers may struggle to find implications for program design in Rosenfield's book, so I have explained his ideas at some length. Edelman's argument is more accessible to AI researchers because he draws on some familiar sources, referring to Lakoff throughout the book, and explicitly discussing models of representation and learning. Nevertheless, the neurobiological argument is intricate and is by and large unfamiliar to AI researchers. Although the theory of Neural Darwinism has been reviewed for this audience [48,49], I present the ideas again to provide an alternative synthesis that includes Edelman's earlier work on topobiology and makes connections to the broader themes of cognitive science that Edelman now wishes to emphasize.

The title *Bright Air, Brilliant Fire* comes from Empedocles, "a physician, poet, and an early materialist philosopher of mind" (p. xvi) in the sixth century B.C., who suggested that perception can be understood in terms of material entities. Edelman believes that understanding the particular material properties of the actual "matter underlying our minds—neurons, their connections, and their patterns" (p. 1) is essential for understanding consciousness and building intelligent machines, because the brain works unlike any machine we have ever built. Of special interest is how Edelman relates his understanding of sensorimotor coordination to Lakoff's analysis of concepts as embodied processes.

The book is organized into four parts: (1) "Problems" with current models of the mind; (2) "Origins" of new approaches based on evolution and developmental biology; (3) "Proposals" for neurobiological models of memory, consciousness, and language; and (4) "Harmonies" or "fruitful interactions that a science of mind must have with philosophy, medicine, and physics" (p. 153). The book concludes with a forty-page postscript, "Mind without Biology", criticizing objectivism, mechanical functionalism, and formal approaches to language. I will cover these ideas in the same order: (1) biological mechanisms that are potentially relevant, and perhaps crucial, to producing artificial intelligence; (2) a summary of the theory of Neural Darwinism; (3) how consciousness arises through these mechanisms; and (4) the synthesis of these ideas in the Darwin III robot. In Section 4, I elaborate on the idea of pre-linguistic coordination, which reveals the limitations of stored linguistic schema mechanisms. In Section 5, I argue for preserving functionalism as a modeling technique, while accepting Edelman's view that it be rejected as a theory of the mind.

3.1. The matter of the mind: biological mechanisms

> If you consider these extraordinary brain properties in conjunction with the dilemmas created by the machine or the computer view of the mind, it is fair to say that we have a scientific crisis .... For a possible way out, let us look to biology itself, rather than to physics, mathematics, or computer science. (p. 69)

Edelman believes that neuroscience now allows us to begin "connecting up what we know about our minds to what we are beginning to know about our brains" (p. 5). His analysis combines an alternative epistemology, which he calls "anti-cognitivist", with biological mechanisms he calls "value-based selectionism" and "Neural Darwinism".

Cognitivism is the view that reasoning is based solely on manipulation of semantic representations. Cognitivism is based on *objectivism* ("that an unequivocal description of reality can be given by science") and *classical categories* (that objects and events can be "defined by sets of singly necessary and jointly sufficient conditions") (p. 14). This conception is manifest in expert systems, for example, or any cognitive model that supposes that human memory consists only of stored linguistic descriptions (e.g., scripts, frames, rules, grammars). Echoing many similar analyses (e.g., [15,29,54]), Edelman characterizes these computer programs as "axiomatic systems" because they contain the designer's symbolic categories and rules of combination, from which all the program's subsequent world models and sensorimotor procedures will be derived. Paralleling the claims of many other theorists, from Collingwood and Dewey to Garfinkel and Bateson, he asserts that such linguistic models "... are social constructions that are the *results* of thought, not the basis of thought" (p. 153). He draws a basic distinction between what people do or experience and their linguistic descriptions (names, laws, scripts):

> Laws do not and cannot exhaust experience or replace history or the events that occur in the actual courses of individual lives. Events are denser than any possible scientific description. They are also microscopically indeterminate, and, given our theory, they are even to some extent macroscopically so. (pp. 162–163)

This distinction between practice and theory dominates anthropological theory [51].

Although science cannot exhaustively describe particular, individual experiences, it can properly study the *constraints* on experience. Edelman focuses on biological constraints. He claims that the *biological organization of matter in the brain produces kinds of physical processes that have not been replicated in computers.* "By taking the position of a biologically-based epistemology, we are in some sense realists" (recognizing the inherent "density" of objects and events) "and also sophisticated materialists" (p. 161) (holding that thought, will, etc. are produced by physical systems, but emphasizing that not all mechanisms have the same capabilities).

Edelman believes that cognitivism produced "a scientific deviation as great as that of the behaviorism it attempted to supplant" (p. 14) in assuming that neurobiological processes have no properties that computers don't already replicate (e.g., assembling, matching, and storing symbol structures). This assumption limits what current computers can do, as manifest in: the symbol grounding problem, combinatorial search, inflexibilities of a rule-bound mechanism, and inefficient real-time coordination. The most "egregious" category mistake is "the notion that the whole enterprise [of AI] can proceed by studying behavior, mental performance and competence, and language under the assumptions of functionalism without first understanding the underlying biology" (p. 15). By this account, Newell's [32] attempts to relate psychological data to biological constraints (not cited by Edelman) are inadequate because the "bands" of *Unified Theories of Cognition* misconstrue the interpenetration of neural and environmental processes (Section 2). Putting "mind back into nature" requires considering "how it got there in the first place.... [We] must heed what we have learned from the theory of evolution".

Edelman proceeds to summarize the basic developmental neurobiology of the brain, "the most complicated material object in the known universe" (p. 17). Development is epigenetic, meaning that the network and topology of neural connections is not prespecified genetically in detail, but develops in the embryo through competitive neural activity. Surprisingly, cells move and interact: "in some regions of the developing nervous system up to 70 percent of the neurons die before the structure of that region is completed!" (p. 25).[1] The brain is not organized like conventionally manufactured hardware: the wiring is highly variable and borders of neural maps change

[1] Formation of primary connections between cells may continue in the development of the immature organism, intermixed with formation of secondary repertoires (neuronal groups). The nature of synaptic and neuronal group selection changes after adolescence, affecting acquisition of new languages in adults (explaining Rosenfield's analysis of language evolution over multiple generations).

over time. Individual neurons cannot carry information in the sense that electronic devices carry information, because there is no predetermination of what specific connections and maps mean:

> Nervous system behavior is to some extent self-generated in loops; brain activity leads to movement, which leads to further sensation and perception and still further movement. The layers and loops ... are dynamic; they continually change. (p. 29)

As previously mentioned, Dewey [17] emphasized that neural activations arise as complete *circuits*, within already existing coordinations (sequences of neural activations over time), not isolated paths between peripheral subsystems. Carrying the idea further, Edelman states that "there is no such thing as software involved in the operations of brains" (p. 30). As we will discuss later (Sections 3.2 and 4), this means that each new perceptual categorization, conceptualization, and sensory-motor coordination brings "hardware" components together in new ways, modifying the population of physical elements available for future activation and recombination. Crucially, this physical rearrangement of the brain is not produced by a software compilation process (translating from linguistic descriptions) or isomorphic to linguistic names and semantic manipulations (our conventional idea of software). Different structures can produce the same result, so "there is macroscopic indeterminacy ... the strong psychological determinism proposed by Freud does not hold" (pp. 169–170).

Edelman observes that only biological entities have intentions, and asks, what kind of morphology provides a minimum basis for mental processes, and "when did it emerge in evolutionary time"? (p. 33) . How did the brain arise by natural selection? By better understanding the development of hominid behavior in groups and the development of language, we can better characterize the function and development of mental processes, and hence understand how morphology was selected. Given the 99% genetic similarity between humans and chimpanzees, we would do well to understand the nature, function, and evolution of the differences. Edelman seeks to uncover the distinct *physical capabilities* that separate animals from other life and humans from other primates. What hardware organizations make language and consciousness possible?

3.2. Neural Darwinism: the sciences of recognition

Edelman received the Nobel Prize in 1972 for his model of the recognition processes of the immune system. Recognition of bacteria is based on competitive selection in a population of antibodies. This process has several intriguing properties (p. 78):

(1) there is more than one way to recognize successfully any particular shape;
(2) no two people have identical antibodies;
(3) the system exhibits a form of memory at the cellular level (prior to antibody reproduction).

Edelman extends this theory to a more general "science of recognition":

> By "recognition," I mean the continual adaptive matching or fitting of elements in one physical domain to novelty occurring in elements of another, more or less independent physical domain, a matching that occurs without prior instruction. ... [T]here is no explicit information transfer between the environment and organisms that causes the population to change and increase its fitness. (p. 74) [2]

By analogy, mental categories, coordinations, and conceptualizations are like a population of neural maps constituting a "species". There is a common *selectional mechanism* by which the organism "recognizes" an offending bacteria, as well as "recognizes" an experienced situation:

> Memory is a process that emerged only when life and evolution occurred and gave rise to the systems described by the sciences of recognition.... [I]t describes aspects of heredity, immune responses, reflex learning, true learning following perceptual categorization, and the various forms of consciousness.... What they have in common is *relative stability of structure under selective mapping events.* (pp. 203–204)
>
> The species concept arising from ... population thinking is central to all ideas of categorization. Species are not 'natural kinds'; their definition is relative, they are not homogeneous, they have no prior necessary condition for their establishment, and they have no clear boundaries. (p. 239)

The theory explains "how multiple maps lead to integrated responses, and how they lead to generalizations of perceptual responses, *even in the absence of language*" (p. 82, emphasis added).

Edelman's theory of neuronal group selection (TNGS) has several components:

[2] Here Edelman follows von Foerster's [53] usage, suggesting that the term "information" be reserved for categories constructed by an organism in segmenting and classifying signals. Maturana goes a step further, insisting that in labeling phenomena as "signals" an observer is partitioning a single interactive process into "inside" and "outside" components and events. This is an important aspect of scientific study, but should not suggest that the analytic categories have existence apart from the observer's ontology and purposes (for discussion, see [14,54]).

(1) how the structure of the brain develops in the embryo and during early life (topobiology);
(2) a theory of recognition and memory rooted in "population thinking" (Darwinism); and
(3) a detailed model of classification and neural map selection (Neural Darwinism).

Topobiology "refers to the fact that many of the transactions between one cell and another leading to shape are place dependent" (p. 57). This theory partially accounts for the nature and evolution of three-dimensional functional forms in the brain. Movement of cells in epigenesis is a statistical matter (p. 60), leading identical twins to have different brain structures. Special signaling processes account for formation of sensory maps during infancy (and in some respects through adolescence). The intricacy of timing and placement of forms helps explain how great functional variation can occur; this diversity is "one of the most important features of morphology that gives rise to mind" (p. 64). Diversity is important because it lays the foundation for recognition and coordination based exclusively on selection within a population of (sometimes redundant) connections.

Population thinking is a characteristically biological mode of thought "not present or even required in other sciences" (p. 73). It emphasizes the importance of diversity—not merely evolutionary change, but selection from a wide variety of options. "Population thinking states that evolution produces classes of living forms from the bottom up by gradual selective processes over eons of time" (p. 73). Applied to populations of neuronal groups, there are three tenets:

- *developmental selection*, through epigenetic processes already mentioned,
- *experiential selection*, the creation of a secondary level repertoire, called neuronal groups, through selective strengthening and weakening of the neural connections, and
- *reentry*, which links two maps bi-directionally through "parallel selection and correlation of the maps' neuronal groups" (p. 84).

The levels of nested components involved in categorization are: neural cells, neuronal groups, neural maps, classification couples, and global maps. I summarize these components in the following two subsections.

3.2.1. Neuronal groups and classification

Neuronal groups are collections of neural cells that fire and oscillate together (p. 95). Neuronal groups are the units of selection in the development of new functioning circuits (pp. 85–86). By analogy to organisms in a species and lymphocytes, neuronal groups are individuals (Table 1). Reactivation of a neuronal group corresponds to selection of individuals in

Table 1
Neuronal group selection viewed according evolutionary Darwinism.

Species	Functionally segregated map, responding to local features and participating in classification couples with other maps.
Population	Map composed of neuronal groups.
Individual	Neuronal group.

a species.[3] Although one might suppose individual synapses or neurons to correspond to individuals in a population, individual neurons are in general always selected within a group and only influence other neurons through groups: Each neural cell "receives inputs from cells in its own group, from cells in other groups, and from extrinsic sources" (p. 88).[4] The existence of neuronal groups is controversial, but has been experimentally demonstrated (pp. 94–95).

A *neural map* is composed of neuronal groups. Two functionally different neural maps connected by reentry form a *classification couple*:

> Each map *independently* receives signals from other brain maps or from the world.... [F]unctions and activities in one map are connected and correlated with those in another map.... One set of inputs could be, for example, from vision, and the other from touch. (p. 87)

Edelman doesn't relate neuronal selection as clearly to species evolution as we might expect for a popularized treatment. I attempt here and in Table 1 to make the connections more explicit. First, a significant number of non-identical neuronal groups can function similarly within maps (responding to the same inputs), a fundamental property of TNGS called *degeneracy* [21, p. 6]. This roughly corresponds to different individuals in a species having different genotypes, but selected within an environment for similar functional characteristics. Apparently, a population of neuronal

[3] Note that the ideas of "mating" and "reproduction" are not essential parts of the more general ideas of population thinking and recognition. Apparently, the reactivation of a neuronal group corresponds to reproduction of a new individual with "inherited" relations from its activation within previous maps. Changes in genotype of individuals in a species correspond to changes in strength of synaptic connections of neuronal groups within a map (p. 94). A simple evolutionary analogy might suggest viewing an individual as an *instance* of a species. Instead, we view a species as a coherent collection of *interacting* individuals (here a map of neuronal groups). Thus, the connections define the population. Furthermore, selection occurs on multiple levels of form—neuronal groups, maps, and maps of maps.

[4] Formation of synaptic connections (primary repertoire) and neuronal groups (secondary repertoire) can be intermixed (p. 85). The extraordinary, three-fold increase in human brain size after birth [30, p. 159] may be related to the formation of reentrant loops between the conceptual cortex and perceptual categorization, enabling primary consciousness (Fig. 1).

groups becomes a "species" when it becomes functionally distinct from other populations. This occurs when maps interact during the organism's behavior. In effect, the "environment" for a map consists of other active maps. Excitatory and inhibitory interactions between maps correspond to interspecies interactions at the level of competitive and symbiotic relations in the environment. Neural maps effectively define each other's populations by activation relations between their neuronal groups. Reentry (bi-directional activation between *populations* of neuronal groups)[5] provides the means for map interaction and reactivation during organism behavior. Reentry explains how "brain areas that emerge in evolution coordinate with each other to yield new functions" (p. 85) during an individual organism's lifetime. Specifically, local maps can be *reused without copying* by selection of additional reentry links to form new classification couples (with specialized interactions between their neuronal groups). Edelman concludes that reentry thus provides "the main basis for a bridge between physiology and psychology" (p. 85).

3.2.2. Coordinating categorizations by global maps: sequences and concepts

Another level of organization is required to dynamically coordinate categorizations to ongoing sensorimotor behavior: "A global mapping is a dynamic structure containing multiple reentrant local maps (both motor and sensory) that are able to interact with non-mapped parts of the brain" (p. 89). Selection continually occurs within local maps of a global map, making connections to motor behavior, new sensory samplings, and successive reentry events, allowing new categorizations to emerge:

> Categorization does not occur according to a computerlike program in a sensory area which then executes a program to give a particular motor output. Instead, sensorimotor activity over the whole mapping *selects* neuronal groups that give appropriate output or behavior, resulting in categorization. (pp. 89–90)

Appropriateness is determined by internal criteria of value that constrain the domains in which categorization occurs, exhibited most fundamentally in regulation of bodily functions (respiratory, feeding, sex, etc.):

> The thalamocortical system ... evolved to receive signals from sensory receptor sheets and to give signals to voluntary muscles. ... [I]ts main structure, the cerebral cortex is arranged in a set of maps ... as highly connected, layered local structures with massively reentrant connections.... [T]he cortex is concerned with the categorization of the world and the limbic-brain system

[5] See Reeke et al. [40] for further comparison of reentry to recursion.

> is concerned with value. ... [L]earning may be seen as the means by which categorization occurs on a background of value ... (pp. 117–118)

Categorization is therefore relational, occurring within, and in some sense bound to, an active, *ongoing coordinated sequence* of sensory and motor behavior: "The physical movements of an animal drive its perceptual categorization ... " (p. 167). Crucially, global maps themselves rearrange, collapse, or are replaced through perturbations at different levels (p. 91).

Memory "results from a process of continual *recategorization.* By its nature, memory is procedural and involves continual motor activity" (p. 102). Hence, memory is not a place or identified with the low-level mechanisms of synaptic reactivation; and certainly memory is not a coded representation of objects in the world (p. 238). Rather, "memory is a system property" (p. 102) involving not only categorization of sensory-motor activations, but categorizations of *sequences* of neural activations:

> The brain contains structures such as the cerebellum, the basal ganglia, and the hippocampus that are concerned with timing, succession in movement, and the establishment of memory. They are closely connected with the cerebral cortex as it carries out categorization and correlation of the kind performed by global mappings.... (p. 105)

> The brain ... has no replicative memory. It is historical and value driven. It forms categories by internal criteria and by constraints acting at many scales. (p. 152)

Following Lakoff's analysis, Edelman distinguishes between concepts and linguistic symbols. *Concept* refers "to a capability that appears in evolution prior to the acquisition of linguistic primitives.... Unlike elements of speech, however, concepts are *not* conventional or arbitrary, do *not* require linkage to a speech community to develop, and do *not* depend on sequential presentation" (p. 108). Concepts are categorizations of internal categorizing:

> [I]n forming concepts, the brain constructs maps of its *own* activities.... [These maps] categorize parts of past global mappings according to modality, the presence or absence of movement, and the presence or absence of relationships between perceptual categorizations.... They must represent *a mapping of types of maps.* Instead, they must be able to activate or reconstruct portions of *past* activities of global mappings of different types.... They must also be able to recombine or compare them. This means that special reentrant connections from these higher-order cortical areas

> to other cortical areas and to the hippocampus and basal ganglia must exist to carry out concepts. (p. 109)

Thus, *intentional behavior* involves sensory-motor sequencing influenced in a top-down manner by conceptual reactivation and construction: "because concept formation is based on the central triad of perceptual categorization, memory, and learning, it is, by its very nature, intentional" (p. 110).

3.3. Consciousness

This brings us to consciousness, which Edelman characterizes on two levels: *primary consciousness*, found in some animals such as dogs, and *higher-order consciousness*, found in humans and to some degree in other primates:

> Primary consciousness is the state of being mentally aware of things in the world—of having mental images in the present. But it is not accompanied by any sense of a person with a past and a future.... In contrast, higher-order consciousness involves the recognition by a thinking subject of his or her own acts or affections. It embodies a model of the personal, and of the past and the future as well as the present. It exhibits direct awareness—the noninferential or immediate awareness of mental episodes without the involvement of sense organs or receptors. It is what we humans have in addition to primary consciousness. We are conscious of being conscious. (p. 112)

In effect, Edelman claims that a special kind of physical link between conceptual and perceptual categorization enables being tacitly aware of ourselves in relation to what we have done before or imagined will occur. This self-reference, the dynamic coordination of action and attention described by Rosenfield, involves conceptualization of internal experience in a manner that is "direct", rather than mediated by deliberation. Nevertheless, linguistic naming and inference in our previous activity over time plays a key role in self-conceptualization. For example, according to this theory, comprehending a set of instructions involves conceptualization that enables oriented action in the future, without consulting the instructions. Of course, linguistic representation (naming, telling stories, giving explanations) is essential for dealing with an inability to coordinate activity by primary consciousness alone ("impasses" described by Bartlett [5] and "breakdowns" described by Winograd and Flores [54]). For example, in following through a previously comprehended plan, we may become aware that we don't know what to do next because we sense that a situation is unfamiliar.

3.3.1. Primary consciousness: categorizing qualia into a scene

Paralleling Dennett's [15] analysis, Edelman suggests that reports of subjective experience can be correlated and used as a basis for the scientific study of consciousness.

> Qualia constitute the collection of personal or subjective experiences, feelings, and sensations that accompany awareness.... For example, the "redness" of a red object is a quale. (p. 114)
>
> [Q]ualia may be usefully viewed as forms of higher-order categorization, as relations reportable to the self.... (p. 116)
>
> [I]n some animal species with cortical systems, the categorizations of separate causally unconnected parts of the world can be correlated and bound into a *scene* ... a spatiotemporally ordered set of categorizations of familiar and unfamiliar events. ... [T]he ability to create a scene ... led to the emergence of primary consciousness. (p. 118)

Three evolved functions are sufficient for primary consciousness:

- a cortical system linking conceptual functions to the limbic system;
- a "value-category" memory, allowing "conceptual responses to occur in terms of the *mutual* interactions of the thalamocortical and limbic-brain stem systems" (p. 119); and
- "continual reentrant signaling between the value-category memory and the ongoing global mappings that are concerned with perceptual categorization in real time".

Linking perceptual events into a *scene* constitutes "a conceptual categorization of concurrent perceptions". This occurs *before* the independent perceptual signals contribute to independent memory of each modality (p. 119). As an adaptive way of directing attention, this mechanism accounts for *how a sense of similarity arises* prior to articulation of categorical features in metaphorical reasoning [44]. Current value-free perceptual categorization is interacting with the value-dominated conceptual memory before perceptual events and subsequent linguistic theorizing modify conceptual memory. Edelman calls this effect "the remembered present" (p. 120):

> An animal with primary consciousness sees the room the way a beam of light illuminates it.... In all likelihood, most animals and some birds may have it .... [W]e can be fairly sure that animals without a cortex or its equivalent lack it .... (pp. 122–123)

Primary consciousness in itself does not constitute *awareness of having a long-term memory* or ability to plan based on it:

> Perceptual categorization ... is nonconscious and can be carried out by classification couples.... It treats *signals from the outside world.* By contrast, conceptual categorization works from within the brain, requires perceptual categorization and memory, and treats *the activities of portions of global mappings* as its substrate. Connecting the two kinds of categorization with an additional reentrant path for each sensory modality (that is, in addition to the path that allows conceptual learning to take place) gives rise in primary consciousness to a correlated scene, or "image." [A]n animal with primary consciousness alone is strongly tied to the succession of events in real time. (p. 125)

An animal with only primary consciousness can have long-term memories and act upon them, but "cannot, in general, be aware of that memory or plan an extended future for itself based on that memory" (p. 122). Additional categorization loops tied to linguistic actions enable us to transcend the immediacy of primary consciousness. As Bartlett put it:

> If only the organism could hit upon a way of turning round upon its own 'schemata' and making them the objects of its reactions .... It would be the case that the organism would say, if it were able to express itself: "This and this and this must have occurred, in order that my present state should be what it is". And, in fact, I believe this is precisely and accurately just what does happen in by far the greatest number of instances of remembering .... [5, p. 202]

Primary consciousness involves *internal* criteria to "determine the salience of patterns". Higher-order consciousness "adds socially constructed selfhood", further freeing the individual from "the constraints of an immediate present" (p. 133).

3.3.2. Higher-order consciousness: linguistically modeling past and future

Higher-order consciousness involves language for modeling the relation of the self to the world and interactions with other members of the species. But "a model of self–nonself interaction probably had to emerge prior to a true speech", as is indicated by chimpanzee behavior (p. 126). By such conceptualization, an animal can experience higher-order consciousness without language, but not represent what it means or employ it to reason *about* problems. Language involves naming, telling stories about the past, constructing causal theories, modeling designs for new artifacts, and comparing plans for future actions. In humans, "consciousness of being conscious" involves linguistically representing "a true self (or social self) acting on an environ-

Table 2
Relation of primary and higher-order consciousness.

Consciousness	Morphological requirements	Who experiences it?	Key features
Primary	Cortex reentrant loop connecting value-category memory to current perceptual categorization.	Chimpanzees, probably most mammals and birds	Awareness of directed attention in activity (awareness of intention; basic self-reference).
Higher-order	Broca's and Wernicke's areas; bootstrapping perceptual categorization through linguistic symbolization.	Humans	Awareness of having previous experiences, imagining experiences; conceptualization of self, others, world (awareness of self-reference).

ment and vice versa" (p. 131).[6] Individuals with impaired linguistic ability may have a self-concept that is historically and socially distorted from the perspective of other members of the group (cf. the case of Monsieur A).

Creation of language in the human species required the evolution of (1) cortical areas (named after Broca and Wernicke) to finely coordinate "acoustic, motor, and conceptual areas of the brain by reentrant connections ... [serving] to coordinate the production and categorization of speech" and (2) another layering of categorization, on top of conceptualization, to provide "the more sophisticated sensorimotor ordering that is the basis of true syntax" (p. 127). See Table 2 and Fig. 1.[7]

In a process called "semantic bootstrapping", "the brain must have reentrant structures that allow semantics to emerge *first* (prior to syntax) by relating phonological symbols to concepts" (p. 130). By "phonological sym-

[6] Edelman presents only abstract descriptions of "discriminating qualia", "delaying responses", and "inner events that are recalled". A psychologist or anthropologist would have given at least one example. Is higher-order consciousness necessary to buy groceries? To stalk game? To plant a crop? The evolutionary interactions of language, culture, and consciousness remain obscure. What happens when we comprehend a written plan? Diagnose a patient? Edelman hasn't even begun to relate his model of the brain to existing models of reading, problem solving, and understanding.

[7] Recalling Rosenfield's remark that Edelman should emphasize relations instead of categories, we might remain alert to misleading aspects of such diagrams. In particular, "correlation" occurs as areas of the brain configure each other; particular kinds of categories or memories are not *located* in particular boxes. If we treat boxes uniformly as structural areas of the brain and the lines as reentrant activation links, then classification occurs as a coupled reconfiguration of maps (composed of neuronal groups) within two or more boxes. That is, categorizing physically exists only as a process of co-configuring *multiple areas*, not as stuff stored in some place.

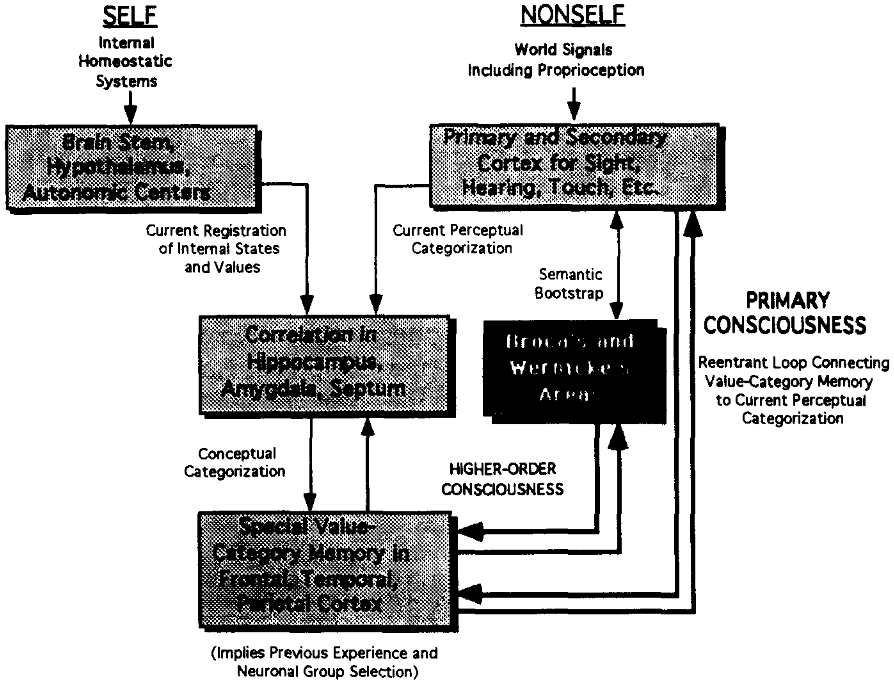
SELF
Internal Homeostatic Systems
NONSELF
World Signals Including Proprioception
Brain Stem, Hypothalamus, Autonomic Centers
Primary and Secondary Cortex for Sight, Hearing, Touch, Etc.
Current Registration of Internal States and Values
Current Perceptual Categorization
Semantic Bootstrap
PRIMARY CONSCIOUSNESS
Reentrant Loop Connecting Value-Category Memory to Current Perceptual Categorization
Correlation in Hippocampus, Amygdala, Septum
Broca's and Wernicke's Areas
Conceptual Categorization
HIGHER-ORDER CONSCIOUSNESS
Special Value-Category Memory in Frontal, Temporal, Parietal Cortex
(Implies Previous Experience and Neuronal Group Selection)

Fig. 1. A scheme for higher-order consciousness. (From Edelman's *Bright Air, Brilliant Fire*, p. 132.)

bols" Edelman apparently means words, viewed as acoustic categorizations. He goes on to say, "When a sufficiently large lexicon is collected, the conceptual areas of the brain categorize the *order* of speech elements". Thus, syntactic correspondences are generated, "not from preexisting rules, but by treating rules *developing in memory* as objects for conceptual manipulation". This is a memory for actual speaking coordinations, not a memory of stored grammar expressions. Edelman doesn't clearly say what he means by conceptual manipulation, but it presumably involves recategorization of previous symbol sequences, as well as categorization of the relation of concepts to symbol sequences:

> The addition of a special symbolic memory [the lexicon of words and phrases] connected to preexisting conceptual centers results in the ability to elaborate, refine, connect, create, and remember great numbers of new concepts.... Meaning arises from the interaction of value-category memory with the *combined* activity of the conceptual areas and speech areas. (p. 130)

Thus, there are stages of intention, reference, awareness, and control: Conceptualization enables an animal to exhibit intention. Primary consciousness involves *awareness of intention*, relating the self to ongoing events. Through categorizations of scenes involving intentional acts of self and others, animals with primary consciousness can exhibit an understanding of *reference* (e.g., a dog seeing a ball knowing that a game is beginning). Intentional acts are imagined, modeled, and controlled through linguistic actions over time in higher-order consciousness. With language, reference becomes symbolic. *Self-reference* begins as value-oriented categorization. When concepts of the self, the past, and the future relate conceptual-symbolic models produced in speech to ongoing perceptual experience, we become aware of self-reference and consciously direct it (e.g., Monsieur A's statement about the need to say things to himself in order to remember them). Reentrant loops give us the ability to project visual, verbal, and emotional experiences; we can attentively enact previously imagined actions—"as if one piece of spacetime could slip and map onto another piece" (p. 169). The problem of coordinating awareness of doing, talking, and visualizing so as to be "consciously unconscious" is a well-known problem in sports [24]. "We live on several levels at once" (p. 150).

> [I]n human beings, primary consciousness and higher-order consciousness coexist, and they each have different relations to time. The sense of "time past" in higher-order consciousness is a *conceptual* matter, having to do with previous orderings of categories in relation to an immediate present driven by primary consciousness. Higher-order consciousness is based not on ongoing experience, as is primary consciousness, but on the ability to model the past and the future. (pp. 167–168)

Once a socially-constructed self arises as a result of higher consciousness, the self becomes necessary "to link one mental image to the next in order to appreciate the workings of primary consciousness" (p. 124).

> Qualia, individual to each of us, are recategorizations by higher-order consciousness of value-laden perceptual relations in each sensory modality or their conceptual combinations with each other .... [They] are increasingly refined by language ... [A] world is developed that requires naming and intending. (p. 136)

Thus, reentrancy and bootstrapping from symbolic and conceptual memories becomes a necessary part of ongoing perceptual categorization. "Consciousness appeared as a result of natural selection. The mind depends on consciousness for its existence and functioning" (p. 149).

Edelman extends his model to broadly explain how neural dysfunctions lead to the kinds of behavior discussed by Rosenfield. He underscores that "All mental diseases are based on physical changes" (p. 178). In particular, he believes that Freud's explanations are limited by inadequately characterizing biological processes:

> "Neurological disease" refers to disruptions of sight, movement, and so forth, and is the result of alterations in the regions of the brain involved in these functions. "Psychiatric disease" refers to alterations in categorization, mental activity, qualia, and so forth, in which responses are symbolically deviant or in which "reality testing" is compromised. (p. 181)

For example, schizophrenia may be a "disease of reentry" produced by a "disabling of communications between reentrant maps ..." (p. 184) resulting in overdomination of a perceptual mode (e.g., producing hallucinations), difficulty coordinating the organs of succession, or discoordination between "the lexicon, conceptual centers, and those that mediate imagery" (p. 185). Although not as dramatic as the effects of psychosis, the discomfort experienced by the patients discussed by Rosenfield apparently follows from their impaired ability to reestablish such relationships within a conscious "scene": "the patient's overall response is still an attempt at adaptation, at reintegration" (p. 185).

3.4. Design of Darwin III: synthetic neural systems

Thanks to Reeke et al. [40], Edelman's theories are being tested by development of computer models. Edelman strongly supports the constructive approach of AI: "the only way we may be able to integrate our knowledge of the brain effectively, given all its levels, is by synthesizing artifacts" (p. 188). He proposes the term "noetics" for devices that "act on their environment by selectional means and by categorization on value" (p. 192), in contrast with devices that adapt only within fixed, predesigned constraints (cybernetics) or programmed devices (robotics).

Darwin III[8] is a "recognition automata that performs as a global mapping" (p. 92) that coordinates vision with a simulated tactile arm in a simulated environment. It is capable of "correlating a scene" by reentry between value-category memory and perceptual categorizations. Values are built in (e.g., light is better than darkness), but the resulting categorizations are all internally developed. The system consists of 50 maps, containing 50,000 cells and over 620,000 synaptic junctions [40, p. 608]. This system rests on the model of "reentrant cortical integration" (RCI) which has been tested with much larger networks (129 maps, 220,000 cells, and 8.5 million connections) that simulate visual illusions and the detection of structure from motion in the monkey's visual cortex.

The statistical, stochastic nature of selection is common to many connectionism models. It was mentioned by Bateson [6] in his own discussion of

[8] This program shouldn't be confused with Calvin's "Darwin Machine" [9, p. 372], which was proposed five years after the initial work by Reeke and Edelman.

parallels between the evolution of biological phenotypes and the development of ideas. Edelman's model probes deeper by specifying how neural nets are *grown*, not merely selected, and how learning is based on internal value. Neural Darwinism can be contrasted with other connectionist approaches in these aspects:

- the influence of epigenetic and infant development as the source of variability;
- degenerate (redundant) populations of preferred maps for recognition;
- selection that is not merely eliminative (the rich get richer), but maintains variability;
- details concerning global mapping, reentrancy, sensorimotor maps, generalization, classification couples.

We can also apply Pagels' criteria [35, pp. 140–141] for comparing connectionists' models. First, like connectionist models, Darwin III is *not neurally realistic* and arguably lacks massive parallelism. But unlike most connectionist models, Darwin III is *not constructed by building in words* referring to concepts and things in the world that it will learn about [39]. Finally, Darwin III is *based on a series of principles* involving evolution, selectionism, development, non-encoding nature of representations, and a distinction between concepts and symbols.

NOMAD is a robotic implementation of Darwin III, claimed to be "the first nonliving thing capable of 'learning' in the biological sense of the word" (p. 193). But Edelman demurs of replicating the capabilities of the brain. Building a device capable of primary consciousness will require simulating

> ... a brain system capable of concepts and thus of the *reconstruction* of portions of global mappings.... [A]rtifacts with higher-order consciousness would have to have language and the equivalent of behavior in a speech community.... [T]he practical problems ... are so far out of reach that we needn't concern ourselves with them now. (p. 194)

4. Pre-linguistic coordination

Edelman's and Rosenfield's discussions of non-linguistic coordinations provide a way of understanding the claim that knowledge doesn't consist of stored representations and linguistic programs. Even if the reader doesn't buy their argument that there are no stored linguistic structures, the discussion of coordination reveals the adaptability they believe that programs lack.

To understand why stored linguistic schema models poorly capture the flexibility of human behavior, Rosenfield makes an analogy between posture and speech:

> [There is no] dictionary of all the words I know stored in my brain, waiting for me to use them. I create my language, and my sense of myself, more dynamically, just as I move around bodily in space. My sense of "posture" is not stored in my brain, but, rather, the ability to create one posture from another is, the ability to establish relations. And the senses of self and speech, like posture, are constantly evolving structures; what I just said determines, in part, what I will say. Just as one posture gives rise to another and one sentence gives rise to another, one expression of my personality gives rise to another.
>
> Memory, too, comprises the acquired habits and abilities for organizing postures and sentences—for establishing relations. (p. 122)

Head, an English neurologist working in the early part of this century and a teacher of Bartlett, introduced the term "schema" in this context. In 1920, he wrote:

> Every recognizable change enters into consciousness already charged with its relation to something that has gone before.... For this combined standard, ... we propose the word "schema" .... Every new posture of movement is recorded on this plastic schema, and the activity of the cortex brings every fresh group of sensations evoked by altered posture into relation with it. (pp. 48–49, quoted by Rosenfield)

Head's notion of a schema is not a linguistic description, but neural and sensory activations, similar to the meaning adopted by Bartlett [5] and more recently Arbib [2].[9] Furthermore, what is organized are the continuous series of dispositions, the changes over time, the relation to what has gone before. As stated by Head, "The unit of consciousness, as far as these factors in sensation are concerned, is not a moment of time, but a 'happening'" (p. 49).[10] Rosenfield nicely summarizes this:

> [A]wareness is change, not the direct perception of stimuli. Conscious images are dynamic relations among a flow of constantly evolving coherent responses, at once different and yet derived from previous responses that are part of an individual's past. (p. 85)

[9] Arbib's work (which isn't cited by Edelman) forms a bridge between neurological models like Neural Darwinism and cognitive theories of vision, planning, and learning. Arbib does an especially good job of reconciling the points of view, where Edelman tends to be dismissive.

[10] History does not record whether Head's "happening", so far in advance of the 1960's American theatrical form by the same name, bears any relation to the "be-in" experienced by Heidegger.

To understand this non-symbolic notion of a schema, consider the movement of limbs in space. The places and orientations of our limbs, eyes, fingers are infinite. Yet, we can symbolically model these relations. We can define points and parameterize space as a coordinate system, thus categorizing the locations of sensory surfaces. By doing this, we can effectively describe human motions, mimic motions in animated simulations, and effectively control robotic behavior. We do all this linguistically, in terms of objects, places, and angles we have defined in our modeling endeavor. The resulting parameterization has some degree of precision determined by the categories and scales we have chosen. The possible space of descriptions, learned behaviors, and control will be bound by the grain size of these representational primitives. For a stable environment with specified goals, a given model may fit satisfactorily. But more refined coordination descriptions will require finer distinctions—changing the representational language. As engineers, we can iterate in this way until we reach a satisfactory model for the purposes at hand.

Now, the claim implicit in Edelman's and Rosenfield's argument about biological function (and indeed, implied by Dewey [17]) is that the human sensorimotor system achieves increasing precision in real time, as part of its activity. Learning to be more precise occurs internally, within an active coordination. Animal behavior clearly shows that such adaptations don't require language. Indeed, there is a higher order of learning in people, involving a sequence of behaviors, in which we represent the world, reflect on the history of what we have done, and plan future actions. In this case, exemplified by the engineer redesigning the robot, the representational language develops in conscious behavior, over time, in cycles of perceiving and acting. Newell and Simon [33, p. 7] called this kind of learning a "second-order effect".

Rosenfield and Edelman insist that learning is also primary and is at this level not limited by linguistic representational primitives. Certainly, a scientist looking inside will see that adaptations are bound by the repertoire of neural maps available for selection and the history of prior activations. But first, the learning does not require *reasoning* about programs, either before or after activity. The bounding is in terms of prior coordinations, not *descriptions* of those coordinations, either in terms of the agent's body parts or places in the world. The claim is that this direct recomposition of prior sensorimotor coordinations, in the form of selection of maps and maps of maps, provides a "run-time" flexibility that executing linguistic circumscriptions of the world does not allow.

Indeed, reflection on prior behavior, learning from failures, and representing the world provide another kind of flexibility that this primary, non-linguistic learning does not allow. Chimps are still in trees; men walk on the moon. But understanding the role of linguistic models requires understanding what can be done without them. Indeed, understanding how models are created and used—how they reorient non-linguistic neurological components (how speaking changes what I will do)—requires acknowledging the existence and nature of this nonlinguistic mechanism that drives animal behavior and still operates inside the human. For example, it is obvious that the dynamic restructuring of posture and speech at a certain grain size bear, for certain kinds of knowledgeable performances, a strong isomorphic mapping to linguistic descriptions, as for example piano playing is directed by a musical score. But in the details, we will find non-linguistically controlled improvisation, bound not by our prior descriptions, but by our prior coordinations. For example, the piano player must sometimes play an error through again slowly to discover what finger is going awry, thus representing the behavior and using this description as a means of controlling future coordinations. How that talk influences new neurological compositions, at a level of neural map selection that was not consciously influenced before, becomes a central issue of neuropsychology.

The machine learning idea of "compiled knowledge" suggests that subconscious processes are merely the execution of previously conscious steps, now compiled into automatic coordinations. Edelman and Rosenfield emphasize that such models ignore the novel, improvised aspect of every behavior. Certainly the model of knowledge compilation has value as an abstract simplification. But it ignores the dynamic mechanism by which sensorimotor systems are structured at run-time with fine relational adjustments that exceed our prior verbal parameterizations. And for animals, such models of learning fail to explain how an animal learns to run through a forest and recognize prey without language at all.

In learning to ski, for example, there is a complex interplay of comprehending an instructor's suggestions, automatically recomposing previous coordinations, and recomposing (recalling) previous ways of describing what is happening. Behavior is coordinated on multiple levels, both linguistic and nonlinguistic, with prior ways of talking, imagined future actions, and attention to new details guiding automatic processes. The important claim is that representing what is happening, as talk to ourselves and others, occurs in our conscious behavior, that it is a manifestation of consciousness, and that it must necessarily be conscious in order to have deliberate, goal-directed effect. Dewey and Ryle's claim that deliberation occurs in our behavior, and not in a hidden way inside, is another way of framing Edelman and Rosenfield's claim that we must understand the structure of consciousness, the progressive flow of making sense of experience, if we wish to understand human cognition.

5. Edelman's view of functionalism

In the appendix to this book, "Mind without Biology: A Critical Postscript", Edelman removes his gloves, and tells us that it is necessary to engage "in a bit of bashing" (p. 211). Evidently, most AI researchers and cognitive scientists have "unknowingly subjected themselves to an intellectual swindle" (p. 229). Despite the many accomplishments of these fields, "an extraordinary misconception of the nature of thought, reasoning, meaning, and of their relationship to perception has developed that threatens to undermine the whole enterprise" (p. 228). What follows is an analysis of "one of the most remarkable misunderstandings in the history of science". Perhaps understandably, some readers have been incensed by this treatment:

> Edelman [22] is one theorist who has tried to put it all together, from the details of neuroanatomy to cognitive psychology to computational models to the most abstruse philosophical controversies. The result is an instructive failure. It shows in great detail just how many different sorts of question must be answered before we can claim to have secured a complete theory of consciousness, but it also shows that no one theorist can appreciate all of the subtleties of the problems addressed by the different fields. Edelman has misconstrued, and then abruptly dismissed, the work of his potential allies, so he has isolated his theory from the sort of sympathetic and informed attention it needs if it is to be saved from its errors and shortcomings. (Dennett, [15, p. 268])

Edelman may go astray in viewing some disciplines outside his own in a stereotyped, monolithic way. Although he would never say "biology believes" or "physics believes" he presents AI and cognitive science as if they were points of view or dogmas, rather than disciplines of study. This error, pointed out by Sloman [47], treats a theory as if it were a field, dismissing the field instead of competing theories within it—a category error. Edelman's position is ironic, given his belief that constructing artificial intelligence systems is possibly the only way to integrate our knowledge of how the brain works.

Edelman's narrow conception of computer science is manifested in his use of the terms "software", "instruction", "computation", "information", "machine", and "computer" itself. For example, he says that it is not meaningful to describe his simulations of artifacts "*as a whole* as a computer (or Turing machine)" (p. 191). Thus, he identifies "computer" with "prespecified effective procedure". This is silly, given that his own system, NOMAD, is built from an N-cube supercomputer. The useful distinctions are the differing architectures, not whether a computer is involved. It is a category error to identify a particular software–hardware architecture as "acting like a computer". Here Edelman speaks like a layperson, as if "the computer" is a theory of cognition.

Unfortunately, this misunderstanding leads Edelman to reject all functional approaches to cognitive modeling. He believes that functionalism characterizes psychological processes in terms of software algorithms, implying that the hardware ("the tissue organization and composition of the brain", p. 220) is irrelevant. From this perspective, functionalism involves promoting a particular kind of hardware architecture, namely that of today's computers, as well as a particular kind of computational model, namely algorithms.

Part of the difficulty is that "functionalism" in cognitive science refers to the idea that principles of operation can be abstractly described and then implemented or emulated in different physical systems (e.g., mental processes are not restricted to systems of organic molecules), as well as the more specific view that *existing computer programs* are isomorphic to the processes and capabilities of human thought (recently stated clearly by Vera and Simon [52]). Within this strong view of Functionalism proper (capital F), proponents vary from claiming that the brain is equivalent to a Turing machine (e.g., Putnam [37]), to saying that "some computer" (not yet designed) with some "computational process" (probably more complex than Soar) will suffice. Johnson-Laird states a version of Functionalism, which Edelman is attacking:

> All theories are abstractions, of course, but there is a more intimate relation between a program modeling the mind and the process that is modeled. Functionalism implies that our understanding of the mind will not be further improved by going beyond the level of mental processes. The functional organization of mental processes can be characterized in terms of effective procedures, since the mind's ability to construct working models is a computational process. If functionalism is correct, it follows not only that scientific theories of mentality can be simulated by computer programs, but also that in principle mentality can be embodied within an appropriately programmed computer: computers can think because thinking is a computational process. [28, pp. 8–9]

This view, sometimes called *mentalism*, is also attacked by Lakoff, the later Putnam [38], Bruner [8], Searle, and many others whom Edelman cites. Some computer scientists find it hard to believe that anybody ever believed in Functionalism, even though it was the everyday working hypothesis that drove the invention of expert systems and cognitive modeling in the 1970s (see [12] for an extended discussion with other quotes from the AI literature).

On the other hand, it is obvious that Edelman accepts the weaker idea of functional descriptions, for he bases his distinction between perceptual, conceptual, and symbolic categorization on Lakoff's analysis of linguistic expressions [29]. Furthermore, Edelman explicitly acknowledges that in focusing on biology, he does not mean that artifacts must be made of organic molecules. When he says that "the close imitation of uniquely biological structures *will* be required" (p. 195), he means that developing artificial intelligence requires understanding the properties of mechanisms that today only exist on earth as biological structures. In this respect, Edelman is just as much a functionalist as Dennett. What he means to say is that certain capabilities may be practically impossible on particular hardware. For example, Pagels [35] argues that it is practically impossible to simulate the brain using a Turing machine, even if it could be so described in principle. AI is a kind of engineering, an effort of practical construction, not of mathematical possibility. Without functional abstractions to guide us, we'd be limited to bottom-up assembling of components to see what develops.

Models of "universal grammar" exemplify how functionalist theories can be reformulated within the biological domain. Arguing against Chomsky's analysis, Edelman claims that Neural Darwinism doesn't postulate "innate genetically specified rules for a universal grammar" (p. 131). But he doesn't consider the possibility that universal grammar may usefully describe (and simplify) the *transformations* that occur as conceptual and symbolic *re-categorizations.* Functional descriptions, as expressed in today's cognitive models, can provide heuristic guidance for interpreting and exploring brain biology.

Contrasting with Edelman's critique, Pagels [35] offers a more accessible analysis of the limitations of cognitive science. Pagels states that "the result of thirty years' work ... [is] brilliantly correct in part, but overall a failure. ... The study of actual brains and actual computers interacting with the world ... is the future of cognitive science" [35, pp. 190–191]. Pagels helps us realize the irony that the quest for a *physical* symbol system so often assumed that the *material processes* of interaction with the world are inconsequential (the Functionalist stance). Thus, mind is disembodied and a timeless, ungrounded mentalism remains.

According to the Physical Symbol System Hypothesis, the material processes of cognition are the data structures, memories, comparators, and read–write operations by which symbols are stored and manipulated. At its heart, Edelman's appeal to biology is a claim that *other kinds of structures* need to be created and recombined, upon which sensorimotor coordination, conceptualization, language, and consciousness will be based. This idea is certainly not new. Dewey [19] argued for biologically-based theories of mind (by which he meant a functional analysis of life experience, akin to

Rosenfield's level of description). Dewey also anticipated problems with exclusively linguistic models of the mind [20].

In conclusion, we might forgive Edelman's "bashing", in view of the fragmentation of views and varying formality of AI research. Edelman can hardly be criticized for adopting the most obvious meanings of the terms prevalent in the literature. In participating in our debate, we can't fault Edelman if he becomes bewildered when we respond, "Not that kind of computer (but one we have yet to invent)" or "Not that kind of memory (but rather one more like what a connectionist hopes to build)". We somehow expect newcomers to be not too critical of what's already on the table, and to sign up instead to the dream.

6. Research conclusions

Despite the different levels of analysis, Rosenfield's and Edelman's books are highly consistent and complementary. Both underscore that perceiving is a form of restructuring previous neural activations, as opposed to matching stored linguistic representations. Both emphasize that developmental stages are grounded in body experience. Both view consciousness as a primary human experience that requires explanation if we are to understand memory and reasoning (but neither cites Dennett). Both believe that theories of cognition must be based on biological arguments about development (but neither cites Dewey or Maturana). Differences lie in the level of discussion: Rosenfield focuses on the nature of consciousness, revealed by clinical data; Edelman provides a broad framework for constructing artificial intelligence, inspired by detailed models of neurological processes. Rosenfield extends the perimeter and depth of a theory of mind; Edelman fills it in.

When studied in detail, Rosenfield's and Edelman's books provide a wealth of new starting points for AI. For example, recently there has been more interest in modeling emotions in AI. These books suggest moving beyond static taxonomies (which are useful early in a scientific effort) to viewing emotions as dynamic, functional, relational experiences. Could the phenomenology of emotional experience be modeled as integral *steps* in sense-making, as Bartlett's model of reminding suggests?

The oddity of Rosenfield's patients, coupled with an evolving architectural model of the brain, often brings to mind questions for further investigation. For example, how did Mr. Baud's inability to remember experience past twenty seconds impair learning new skills or concepts? Cognitive scientists today could easily suggest interesting problems to give Mr. Baud. Similarly, could Gelb and Goldstein's patient, who couldn't understand proverbs or comparisons, make up a story at all? Did she understand causal explanations? Could she describe and rationalize her own behavior? As Rosenfield's book suggests, cognitive neuropsychology is changing. It is time to seek synergy between our disparate models and evidence. As some reviewers of Newell's *Unified Theory of Cognition* suggest, this also entails reconceptualizing what models like Soar describe in relation to the brain (Arbib [3], Pollack [36]).

6.1. Why isn't all reasoning subconscious?

As an example of how cognitive models might be reconceived, consider why a problem solver is aware of intermediate reasoning steps. Nothing in the stored-schema view requires that inference is *consciously* monitored. In Soar, for example, "working memory", where intermediate results come together, corresponds to an agent's awareness, but nothing in the model explains why reasoning is experienced as "phenomenally subjective" [32, p. 434]. Indeed, why isn't all verbal inference subconscious? When we ask a problem solver to think out loud, are we just, like our subject, witnessing ideas spilling over from their more usual, hidden source, like water splashing over a glass? Could verbal thoughts otherwise occur without anyone knowing?

Rosenfield's analysis provides a partial explanation: Representing and inquiry go on *in activity,* that is, in cycles of perceiving and expressing (talking, gesturing, writing) over time. By conjecture, sense-making is necessarily conscious because it involves action and *comprehension of what our acts mean* occurring together. For example, speaking is not just outputting prefabricated linguistic expressions, but a dual process of creating representations in action and comprehending what we are saying. We are "making sense" as we speak—perceiving appropriateness, adjusting, and restating in our activity itself. The process is "dual" because perception, awareness of what we are doing, is integral to every statement. Conscious awareness is not just passive watching, but an active process of sustaining a certain kind of attention that changes the results of inquiry. This analysis suggests specifically that we reconsider "remindings" and other commentary of the subject in experimental problem solving protocols as revealing the *perceptual work* of creating and using representations.

Following Edelman, the rest of the matter, what is going on behind the scenes, is non-linguistic coordination. Conscious acts of fitting—dealing with breakdowns [54]—occur precisely because there is no other place for linguistic representations to be expressed and reflected on, but in our experience itself. This is why we write things down or "talk through" an experience to clarify meanings and implications for future action. Protracted, conscious experience—as in writing a paragraph—is not merely an awareness of elements placed in "working memory", but an active process of recoordinating and recomprehending (reperceiving) what we are doing. In the words of Bartlett, "turning around on our own schemata" is possible precisely because we can recoordinate non-linguistic schemata *in our activity of representing.*

According to Edelman, the articulation process of "building a scene" is reflective at a higher order because of reentrant links between Broca and Wernicke's areas and perceptual categorization (the "semantic bootstrap" of Fig. 1): Our perceptual sense of similarity (reminding) and articulation (naming, history-telling, theorizing) are bound together, so symbolizing actions are driving subsequent perceptions. A sequence of such activity is coordinated by composing a story that accumulates observations and conceptual categories into a coherent sense of what we are trying to do (the scene). In other words, being able to create a story (e.g., the sense-making of a medical diagnostician) is precisely what higher-order consciousness allows. Crucially, human stories are not merely instantiated and assembled from grammars, but are coupled to non-linguistic coordinations grounded in perceptual and motor experience [29]. Hence, consciously-created stories can have an aspect of improvisation and novelty that stored linguistic schema mechanisms do not allow.

This analysis provides an alternative, biologically-grounded perspective on recent arguments about planning and situated action [1,50]. The key idea is that perceiving and acting co-determine each other through reentrant links. Representing occurs in activity, as a means of stepping outside the otherwise automatic process by which neural maps (schemata) are reactivated, composed, and sequenced. Goal-directed, attentive behavior of primary consciousness involves holding active a higher-order organization (global maps) and coordinating the relation to ongoing perceptions (i.e., directed attention). In higher-order consciousness, these global maps are coupled to linguistic descriptions of objects, events, goals, and causal stories.

Robot designers may be impatient with the vagueness of such descriptive theorizing. But it is clear that the clinical evidence and neurobiological mechanisms of Rosenfield and Edelman are adequate to promote further reconsideration of our models of explanation, remembering, and story-telling, including the seminal work of Bartlett.

6.2. Prospective

These two books suggest that subfields within cognitive science are changing and then coming together in new ways. New understanding of neural development and anatomy suggests radical reinterpretation of classic cases of psychological dysfunction. Cognitive neuropsychology is moving away from the stored linguistic schema model of memory. Selectionist models of learning suggest that functional processes can be constructed "in-line", without mediating linguistic descriptions of what the processes do or how the parts fit together. Studies of language and human learning place new primacy on the representations that people see, hear, and manipulate interactively, relegating internal subconscious structures and processes to another level of operation. Interest in modeling animal behavior leads us to reexamine the capabilities of agents without language, and the evolution of consciousness within a social system.

It is tempting to predict that the development of global map architectures, as in Darwin III, will become a dominant approach for neural network research, effectively building on situated cognition critiques of the symbolic approach [13]. However, if it becomes essential to understand the chaotic processes of the brain, as Freeman [23], Pollack [36], and others argue, it is less clear how the researchers who brought us Pengi and Soar will participate in building the next generation of AI machines.

All told, there are probably more pieces here than most researchers can follow or integrate in their work. A good bet is that progress in AI will now depend on more multidisciplinary teams and efforts to bridge these diverse fields. Rosenfield and Edelman make a big leap forwards, showing consciousness to be an evolved activity, grounded in and sustaining an individual's participation in the world as a physical and social personality. With theories like self-reference, population thinking, and selectionism, we pick up Bateson's challenge in *Mind and Nature: A Necessary Unity*, finding the patterns that connect the human world to nature and all of the sciences to each other.

Acknowledgments

The book review editors, Mark Stefik and Steve Smoliar, provided, as usual, a great number of insightful suggestions for clarification and elaboration of my remarks. Funding has been provided in part by a gift from the Xerox Foundation and the National Science Foundation.

References

[1] P.E. Agre, The dynamic structure of everyday life, Ph.D. Dissertation, MIT, Cambridge, MA (1988).

[2] M.A. Arbib, Schema theory, in: S.C. Shapiro, ed., *The Encyclopedia of Artificial Intelligence* (Wiley, New York, 1992).

[3] M.A. Arbib, Book Review of *Unified Theories of Cognition* (Allen Newell), *Artif. Intell.* **59** (1993) 265–283.

[4] J. Bamberger, *The Mind behind the Musical Ear* (Harvard University Press, Cambridge, MA, 1991).

[5] F.C. Bartlett [1932], *Remembering—A Study in Experimental and Social Psychology* (Cambridge University Press, Cambridge, England, 1977).

[6] G. Bateson, *Mind and Nature: A Necessary Unity* (Bantam Books, New York, 1979).

[7] P.L. Berger and T. Luckmann, *The Social Construction of Reality: A Treatise in the Sociology of Knowledge* (Anchor Books, Garden City, NY, 1967).

[8] J. Bruner, *Acts of Meaning* (Harvard University Press, Cambridge, MA, 1990).

[9] W.H. Calvin, *The Cerebral Symphony: Seashore Reflections on the Structure of Consciousness* (Bantam Books, New York, 1990).

[10] W.H. Calvin, Islands in the mind: dynamic subdivisions of association cortex and the emergence of the Darwin Machine, *Neurosci.* 3 (1991) 423–433.

[11] W.J. Clancey, Book Review of *The Invention of Memory* (Israel Rosenfield) *Artif. Intell.* **50** (2) (1991) 241–284.

[12] W.J. Clancey, Representations of knowing: in defense of cognitive apprenticeship, *J. Artif. Intell. Educ.* 3 (2) (1992) 139–168.

[13] W.J. Clancey, Situated action: a neuropsychological interpretation: response to Vera and Simon, *Cogn. Sci.* **17** (1993) 87–116.

[14] P.F. Dell, Understanding Bateson and Maturana: toward a biological foundation for the social sciences, *J. Marital and Family Therapy* **11** (1) (1985) 1–20.

[15] D.C Dennett, *Consciousness Explained* (Little, Brown and Company, Boston, MA, 1992).

[16] D.C. Dennett, Revolution on the mind: reviews of *The Embodied Mind* and *Bright Air, Brilliant Fire, New Scientist* (June 13, 1992) 48–49.

[17] J. Dewey [1896], The reflex arc concept in psychology, *Psychol. Rev.* 3 (1981) 357–370; Reprinted in: J.J. McDermott, ed., *The Philosophy of John Dewey* (University of Chicago Press, Chicago, IL, 1981) 136–148.

[18] J. Dewey, *The Child and the Curriculum* (University of Chicago Press, Chicago, IL, 1902); Reprinted in: J.J. McDermott, ed., *The Philosophy of John Dewey* (University of Chicago Press, Chicago, IL, 1981) 511–523.

[19] J. Dewey, *Logic: The Theory of Inquiry* (Henry Holt & Company, New York, 1938).

[20] J. Dewey and A.F. Bentley, *Knowing and the Known* (Beacon Press, Boston, MA, 1949).

[21] G.M. Edelman, *Neural Darwinism: The Theory of Neuronal Group Selection* (Basic Books, New York, 1987).

[22] G.M. Edelman, *The Remembered Present: A Biological Theory of Consciousness* (Basic Books, New York, 1989).

[23] W.J. Freeman, The physiology of perception, *Sci. Am.* (February 1991) 78–85.

[24] W.T. Gallwey, *The Inner Game of Tennis* (Bantam Books, New York, 1974).

[25] S.J. Gould, Nurturing Nature, in: *An Urchin in the Storm: Essays about Books and Ideas* (W.W. Norton and Company, New York, 1987) 145–154.

[26] A.G. Greenwald, Self and memory, in: G.H. Bower, ed., *The Psychology of Learning and Motivation* **15** (Academic Press, New York, 1981) 201–236.

[27] D.R. Hofstadter, *Gödel, Escher, Bach: An Eternal Golden Braid* (Basic Books, New York, 1979).

[28] P.N. Johnson-Laird, *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness* (Harvard University Press, Cambridge, MA, 1983).

[29] G. Lakoff, *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind* (University of Chicago Press, Chicago, IL, 1987).

[30] R. Leakey and R. Lewin, *Origins Reconsidered: In Search of What Makes Us Human* (Doubleday, New York, 1992).

[31] A.R. Luria, *The Mind of a Mnemonist* (Harvard University Press, Cambridge, MA, 1968).

[32] A. Newell, *Unified Theories of Cognition* (Harvard University Press, Cambridge, MA, 1990).

[33] A. Newell and H.A. Simon, *Human Problem Solving* (Prentice Hall, Englewood Cliffs, NJ, 1972).

[34] R.E. Ornstein, *The Psychology of Consciousness* (Penguin Books, New York, 1972).

[35] H.R. Pagels, *The Dreams of Reason: The Computer and the Rise of the Sciences of Complexity* (Bantam Books, New York, 1988).

[36] J.B. Pollack, On wings of knowledge: a review of Allen Newell's *Unified Theories of Cognition, Artif. Intell.* **59** (1993) 355–369.

[37] H. Putnam, Philosophy and our mental life, in: *Philosophical Papers* **2**: *Mind, Language and Reality* (Cambridge University Press, New York, 1975) 291–303.

[38] H. Putnam, *Representation and Reality* (MIT Press, Cambridge, MA, 1988).

[39] G.N. Reeke and G.M. Edelman, Real brains and artificial intelligence, *Daedalus* **117** (1) (1988) "Artificial Intelligence" issue.

[40] G.N. Reeke, L.H. Finkel, O. Sporns and G.M. Edelman, Synthetic neural modeling: a multilevel approach to the analysis of brain complexity, in: G.M. Edelman, W.E. Gall and W.M. Cowan, eds., *The Neurosciences Institute Publications: Signal and Sense: Local and Global Order in Perceptual Maps* (Wiley, New York, 1990) 607–707 (Chapter 24).

[41] I. Rosenfield, *The Invention of Memory: A New View of the Brain* (Basic Books, New York, 1988).

[42] G. Ryle, *The Concept of Mind* (Barnes & Noble, Inc., New York, 1949).

[43] O. Sacks, *A Leg to Stand On* (Harper Collins Publishers, New York, 1984).

[44] D.A. Schön, Generative metaphor: a perspective on problem-setting in social policy, in: A. Ortony, ed., *Metaphor and Thought* (Cambridge University Press, Cambridge, England, 1979) 254–283.

[45] D.A. Schön, The theory of inquiry: Dewey's legacy to education, presented at the Annual Meeting of the American Educational Association, Boston, MA (1990).

[46] R. Sheldrake, *The Presence of the Past: Morphic Resonance and the Habits of Nature* (Vintage Books, New York, 1988).

[47] A. Sloman, The emperor's real mind: review of Roger Penrose's *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics, Artif. Intell.* **56** (2–3) (1992) 355–396.

[48] S.W. Smoliar, Book Review of *Neural Darwinism: The Theory of Neuronal Group Selection* (Gerald M. Edelman), *Artif. Intell.* **39** (1) (1989) 121–136.

[49] S.W. Smoliar, Book Review of *The Remembered Present: A Biological Theory of Consciousness* (Gerald M. Edelman), *Artif. Intell.* **52** (3) (1991) 295–318.

[50] L.A. Suchman, *Plans and Situated Actions: The Problem of Human-Machine Communication* (Cambridge University Press, Cambridge, England, 1987).

[51] S. Tyler, *The Said and the Unsaid: Mind, Meaning, and Culture* (Academic Press, New York, 1978).

[52] A. Vera and H.A. Simon. Situated action: a symbolic interpretation, *Cogn. Sci.* **17** (1993) 7–48.

[53] H. von Foerster, Epistemology of communication, in: K. Woodward, ed., *The Myths of Information: Technology and Postindustrial Culture* (Coda Press, Madison, WI, 1980).

[54] T. Winograd and F. Flores, *Understanding Computers and Cognition: A New Foundation for Design* (Ablex, Norwood, NJ, 1986).