Artificial Intelligence 50 (1991) 241–284
Elsevier

# Book Review

# Israel Rosenfield, *The Invention of Memory: A New View of the Brain*[*]

William J. Clancey
*Institute for Research on Learning, 2550 Hanover Street, Palo Alto, CA 94304, USA*

Received January 1990

## 1. Introduction

This short, thought-provoking book claims that both machine intelligence research and modern neurobiology are based on faulty interpretations of nineteenth-century clinical studies of human memory. Rosenfield argues against the commonplace view that human memory is a kind of filing cabinet or database, that memories are permanent records, that remembering is retrieving something, that practiced behavior is reexecuting a stored program, and that learning, perceiving, and behaving are separate processes in the brain. Rosenfield's heroes are Freud, Marr, and Edelman. The book is exciting exactly because of this juxtaposition of ideas: Freud's psychiatric interpretations that seek to formalize the origin and effects of emotions (non-symbolic organizers of behavior), Marr's constructivist model of vision (recognition without top-down matching of internal descriptions), and Edelman's developmental approach to connectionism (perception as category learning). Rosenfield challenges the AI researcher to understand the relation between Freud, Marr, and Edelman, arguing that their work supports a radically different, non-representational model of memory—memory for processes of perceiving and behaving rather than memory as descriptions of how the world or behavior appears.

The book is essentially a historical argument for Edelman's work [23]; it grew out of a series of essays originally written for *The New York Review of*

[*] (Basic Books, New York, 1988); x + 229 pages.

*Books*. Edelman's book has already been reviewed here at length (Smoliar [54]), so I will focus on Rosenfield's historical synthesis:

- A reconsideration of nineteenth-century studies of brain-damaged patients suggests new explanations of reading, speaking, and writing dysfunctions, not based on stored memories.
- Freud's work suggests that emotion is not a secondary coloration of memory, but the basis of the constructive process by which we achieve a sense of continuity.
- Marr's work provides a pivotal link between the conventional AI claim that recognition is based on matching internal descriptions of the world and Edelman's neural model of bottom-up perception.

This cast of characters is made all the more interesting when, towards the end of the book, the PDP connectionist approach is lambasted as making the same mistakes as the rest of AI research in its failure to integrate perception, memory, and learning. This highly readable book should be studied by every AI and cognitive science researcher who wishes to understand alternative approaches for designing intelligent machines (e.g., situated automata [36, 59]) or for modeling human behavior (e.g., situated action [1, 56]).

### *1.1. The localization hypothesis*

According to Rosenfield, nineteenth-century interpretations of reading and writing dysfunctions, often caused by brain lesions, assume that memories are fixed. That is, memory is a *place for storing things* where they remain unchanged until they are retrieved. Briefly put, to explain why a patient can speak, but not write, physicians of the day argued for a memory of "permanent traces"—specialized images of things in the world, for example, permanent records of sounds, shapes, colors, and movements. Thus, ability to speak the word "ship", but inability to write it, suggests that the necessary information for speaking and writing is stored separately in the brain. This is called the *localization hypothesis*. Rosenfield demonstrates through a historical survey of the evolution of research that the localization hypothesis has had a major, enduring influence on neurology. Using the work of Freud, Marr, and Edelman, supported by psychological research on perception, Rosenfield argues that this hypothesis is fundamentally wrong.

Rosenfield attacks the view that knowledge consists of stored representations, for example, that we can recognize a table because we retrieve a description of what tables generally look like. Most of AI research is based on this model of memory, epitomized by Quillian's semantic networks, Minsky's frames, and Schank's MOPS, as well as natural language grammars. Knowledge is assumed to consist of stored descriptions of how the world appears (e.g., disease hierarchies, device models) and descriptions of how an agent behaves (e.g., scripts, reasoning strategies). These descriptions are stored as labeled things in memory, so they can be selectively indexed and retrieved, reassembled, and then translated into outward behavior.

But if memory is not a storage place for descriptive structures, "How do I know I'm looking at a table?" In most machine learning research, perception is a peripheral process that feeds objective data to a cognitive matcher; learning involves fine-tuning, composing, and ordering prestored categories (e.g., Norman [42]). The standard AI view is that "There is no perception without prior 'learning'. . ." (p. 7).[1] Rosenfield (after Edelman) argues that perception is not matching internal descriptions of features against sensations. Instead, perceiving is itself a process of categorizing. Perception is not a peripheral process feeding data to an inferential process, but the very act of recognition or understanding itself.

### *1.2. Behaving is coordinating is learning*

Rosenfield claims that the essential problem of categorization goes unaddressed in AI research: ". . . it is unexplained how the images during the initial encounter with information are recognized as worthy of storage" (p. 7). In conventional AI programs, the problem is circumvented by having a program designer build in primitive categories. Rosenfield is thus addressing the well-known problem of how categories get into the brain in the first place. His approach is to overturn the initial assumption that categories are stored *things*. In effect, he claims that categorization occurs at runtime.[2] Put simply, *every new perception or behavior is a generalization*, composed of past perceptions and behaviors (a claim associated with Vygotsky [58]). Current neural organizations are thus related to those constructed in the past, but without an indexing, retrieval, and matching process.

Rosenfield opens the book with a direct fusillade against the idea of stored descriptions, what he calls the Platonic approach:

> This book is about a myth. . . that we can accurately remember people, places, and things because images of them have been imprinted and permanently stored in our brains. (p. 3)
>
> Failure to recall could, therefore, be explained as the loss of a specific image (or center) or as the brain's inability to search its files. . . . (p. 5)

If memory is not a storage place for descriptive forms, then there can be no localization of function at the level of skills like reading and writing. For example, the word "cat" is not stored as a sequence of letter descriptions

[1] Except as noted, all page numbers refer to the book being reviewed.

[2] To be precise, we say "categorizing", not "forming categories", just as Bartlett wrote about "remembering", not "memories" [4].

(c-a-t), whether in one place or in some distributed network; a word, or a concept in general, is not a thing that is put away and retrieved. Reinterpreting the historical data of neurological deficits, Rosenfield postulates instead that functional losses are caused by an inability *to establish correlations*. For example, when attempting to read "cat" as something other than the letters "c-a-t" in sequence, we are correlating a sequence, perceiving, categorizing. To perceive is to compose is to categorize.

In general, "re-collection" involves *coordinated recombination of past processes of perceiving and behaving*. This coordination is organized by (and in some sense subsumes) what the person is currently perceiving and doing. The essence of Rosenfield's critique is that most neuro and cognitive scientists have ignored the nature and role of this ongoing *context* in perceptual categorization: Perception is not peripheral or antecedent to movement, but rather part of a single, coordinated process of behaving, of composing. This means that new processes arise so that they are constituted by processes already occurring. This view unifies intellectual and physical skills, so that at a base level a person is always like a dancer balancing her next steps against the inertia of past movements and her view of where she is going. Throughout the book, examples are given of the role of the ongoing context in correlating present and past behaviors and in coordinating a sequence of coherent movements for some skill (including especially reading and writing).

### *1.3. Similar claims in past research*

Although Rosenfield's claims may first appear outrageous, they are hardly new. In fact, for decades many researchers have made the same claims, supported by psychological experiments. This places me in the ironic position of trying to explain to incredulous AI readers what has already been clearly presented in psychology books and journal articles. None of this work is cited by Rosenfield, except for Bartlett, suggesting that the fragmentation of neurocognitive research is an extensive and serious problem. For this reason, I will quote at some length from the original sources.

A proper historical review might begin with Frederic C. Bartlett's [4] famous work:

> Remembering is not the re-excitation of innumerable fixed, lifeless and fragmentary traces. It is an imaginative reconstruction, or construction, built out of the relation of our attitude towards a whole active mass of organised past reactions or experience, and to a little outstanding detail which commonly appears in image or in language form. [4, p. 213]

> Suppose I am making a stroke in a quick game, such as tennis or cricket. . . . I do not, as a matter fact produce something absolutely

> new, and I never merely repeat something old. The stroke is literally manufactured out of the living visual and postural "schemata" of the moment and their interrelations. [4, p. 202]

> 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. Everytime 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. [4, p. 211]

> I strongly dislike the term "schema". It is at once too definite and sketchy. . . . It suggests some persistent, but fragmentary, "form of arrangement", and it does not indicate what is very essential to the whole notion, that the organised mass results of past changes of position and posture are actively doing something all the time. . . . [4, p. 201]

> Everything in this book has been written from the point of view of a study of the conditions of organic and mental functions, rather than from that of an analysis of mental structure. It was, however, the latter standpoint which developed the traditional principles of associationism. The confusion of the two is responsible for very much unnecessary difficulty in psychological discussion. [4, p. 304]

William James also makes basic distinctions:

> Memory proper, or secondary memory as it might be styled, is. . . *knowledge of an event, or fact,* of which meantime we have not been thinking, *with the additional consciousness that we have thought or experienced it before*. . . . [P]sychical objects (sensations, for example) simply recurring in successive editions will remember each other *on that account* no more than clock-strokes do. No memory is involved in the mere fact of recurrence. [30, p. 252]

In the notes to this page, James wrote, "Faculty view. Ideas not *things* but processes | No reservoir" [30, p. 452].

In spite of this early work, the idea of memory as a storage place took hold and became the basis of the "knowledge is power" movement in AI and cognitive science since the 1960s; it continues in the belief today that common sense knowledge can be collected like so many butterflies. Contemporary psychologists in the past three decades have directly attacked this model:

James J. Gibson:

> The invariance of perception with varying samples of overlapping stimulation may be accounted for by invariant information and by an attunement of the whole retino-neuro-muscular system to invariant information. The development of this attunement, or the education of attention, depends on past experience, but not on the *storage* of past experiences. [25, p. 262]

Jean Piaget:

> I think that human knowledge is essentially active. . . . I find myself opposed to the view of knowledge as a copy, a passive copy of reality. [44, p. 15]
>
> [F]or the genetic epistemologist, knowledge results from continuous construction, since in each act of understanding, some degree of invention is involved. [44, p. 77]

James J. Jenkins:

> [T]he phenomena disclosed by these experiments pose formidable problems for storage theories of memory. [31, p. 792]
>
> [W]e should shun any notion that memory consists of a specific system that operates with one set of rules on one kind of unit. [31, p. 793]
>
> Apart from the belief that the construction of the mind is attributed to the past, he [William James] saw nothing to set memory apart from perception, imagination, comparison, and reasoning. Such a claim is unsettling because it says: *Memory is not a box in a flow diagram*. It is also threatening because it seems to demand an understanding of all "the higher mental processes" at once. Yet, that is what the data in our experiments suggest. To study memory without studying perception is. . . pushing all the difficult problems out of memory into the unknown perceptual domain for someone else to study. [31, p. 794]

John D. Bransford et al.:

> Our purpose is not to deny the importance of *remembering*. . . . But we question the fruitfulness of assuming that a concept of *memory* underlies these events. [C]urrent uses of the term memory involve tacit or explicit assumptions. . . that memory can be broken down into a set of *memories*, that these consist of relatively independent *traces* that are stored in some *location*, that these traces must be *searched for* and *retrieved* in order to produce remembering, and

> that appropriate traces must be "contacted" in order for past experiences to have their effects on subsequent events. [9, p. 431]
>
> [In associative models]. . . the problem of remembering begins where the parsers stop. [9, p. 444]
>
> [W]e believe it unfruitful to separate problems of remembering from problems of comprehending and perceiving. [9, p. 454].

Even researchers sensitive to the complexities of cognition unquestioningly adopt the storage model, to the point of misrepresenting Bartlett:

> As a theory of episodic memory, Bartlett's approach has the interesting implication that general attitudes, undifferentiated as to motor, perceptual, or symbolic content, are stored most faithfully in memory. (Miller and Johnson-Laird [38, p. 150])

Describing Bartlett's model, Miller and Johnson-Laird proceed to talk about "reinstatement" (instantiation of schemas) in every place that Bartlett emphasizes novel construction. This discussion appears in a section titled "memory locations and fields" under the topic of "The organization of memory".

Iran-Nejad writes persuasively about the dominance of the storage metaphor and general blindness to alternatives:

> Counterintuitive as it may seem at first, it is entirely conceivable, however, that the patterning aspect of cognition is a transient functional-phenomenal, rather than a long-term memory structural, organization. [29, p. 115]

Iran-Nejad [28] characterizes cognitive science knowledge models as *intralevel* theories: "They assume that the holistic structures and their constitutive elements are both mental in nature" [28, p. 281]. Following Bartlett he argues that we need an *interlevel* theory to explain how the moment-by-moment group functioning of neural elements creates mental structures (transient, composed processes of correlating, attending, and resolving). Crucially, the causality goes in both directions: ". . . mental structures have a causal influence on the functioning of neuronal elements" [28, p. 283]. By analogy, cognitive science's intralevel theories are like models of a fountain of water that only describe its shape, as if the stable form is produced by an internal template and made out of a fixed set of unchanging parts [28, p. 285].

Bickhard and Richie [7], building on Gibson's theory of perception, outline an interlevel architecture in which mental structures are controlling, nonrepresentational processes embodied as active neural elements ("material processes"):

> From an interactive perspective, however, there is at least one level of emergence *between* the material and the representational: the

> level of interactive control structures. Representation, then, is an emergent functional property of certain forms of goal-directed interactive control structures, which in turn, are emergent properties of certain patterns of material processes. [7, p. 57]

> In present-day information-processing or computational approaches. . . the level of interactive control structures and processes that is properly *between* the material level and the representational level has instead been moved *above* the level of encoded representations, leaving the level of encodings hanging in midair with no grounds for explication. [7, p. 57]

In summary, the idea of memory as stored structures has been criticized and experimentally questioned, but such studies are either ignored or misrepresented in AI and cognitive science research of the past two decades.

### *1.4. Relevance to AI research*

Rosenfield's book is important because it provides an avenue for explaining the "situated cognition" perspective, which has become an important subfield in AI [1, 17, 36]. Situated cognition emphasizes the role of interaction and context in organizing behavior.[3] Rosenschein [47] motivates situated-automata robotics research by criticizing the storage model of memory:

> Since logical sentences are used at the abstract level to express the *content* of knowledge, what could be more natural at the implementation level than to imitate their *form* as well and to think of each distinct fact known by the system as a symbolic assertion stored in the computer's memory?

I believe that the chief contribution of situated cognition research will be to help resolve the learning problem of artificial intelligence by forcing us to abandon the idea that representations are structures stored in memory. In effect, we will be forced to distinguish between representations as they are created and interpreted in perceivable form and the momentary, non-representational constructions that Bickhard and Richie [7] call "interactive control structures".

[3] Jenkins [31] describes the roots of the term in American pragmatism, in the work of William James, C.S. Pierce, and John Dewey. Jenkins calls it *contextualism*. "Contextualism holds that experience consists of events. Events have a quality as a whole. By quality is meant the total meaning of the event. The quality of the event is the resultant of the interaction of the experiencer and the world. . . . For the contextualist, no analysis is 'the complete analysis'." By contrast, "Associationism asserts that there is one correct and final analysis of any psychological event in terms of a set of basic units and their basic relations" [31, p. 787]. By the contextualist view, a knowledge-level description is inherently subjective, relative to the purposes of an observer (cf. Bordieu, [8]).

Simply put, the "perception as coordination" perspective helps us explain how representations are created and given meaning. The relation between knowledge and context is fundamentally changed: Although we may describe knowledge discretely, as a collection of representations that explain an agent's behavior, what we are modeling is a capacity to interact adaptively. This intricate linking of sensation and action cannot be reduced to (replaced by) statements about either the agent or the environment (cf. Winograd and Flores' [61] discussion of Maturana).

Representations, which AI and cognitive science have taken to be the very stuff of inner processing, are instead continuously created in perceptual activity, in an interaction of neural and environmental processes. (To speak is to represent is to perceive, not to translate from something already said privately inside.) Habits, ways of talking, and categories are stable behaviors, not generated from stored descriptions, but continuously reconstructed, albeit strongly biased by previous perceptual-motor compositions. What we have taken in AI to be the inner stuff of cognition—grammars, scripts, strategies—are observer-relative descriptions of patterns of behavior, stable interactions between the agent and his environment which develop over time. To behave according to a pattern is not to be following a template-thing. The pattern description, what we generally call a representation of the agent's knowledge, exists only in the statements, writing, and diagrams of the observer-theoretician [16, 17].

Although Rosenfield intends to make contact with the work of machine intelligence and contrasts it with Marr and Edelman at some length, I don't believe that he is sufficiently familiar with the context sensitivity of goal-driven and machine learning programs to be convincing to most AI readers. Many statements and turns of phrase may appear too loose or ungrounded. If you don't reject claims about context sensitivity outright, you are likely to say, "But that's just what program X can do!" If you believe that MOPS explains the mechanism of reminding (Schank, [51]), if you believe that Bartlett [4] supported the idea of schema-structures, or if you believe that connectionism (Rumelhart and McClelland [48]) shows how knowledge could be distributed in the brain, then you are likely to have trouble reading this book. But you also have a lot to gain.

### *1.5. Outline of this review*

The objective of this review is to bridge the gap between Rosenfield's and the typical AI researcher's perspective. I begin by clarifying what kind of memory Rosenfield is arguing against, attempting to anticipate common misunderstandings. By placing these positive aspects of the discussion first, I am hoping the reader will become sympathetic to the not-stored-structures thesis, and even enthusiastic to learn about Rosenfield's historical analysis in the sections which follow:

- classical neuroscience explains deficiencies in terms of localizable memories (fixed traces), which Rosenfield argues against;
- Freud's work shows how *emotion, a non-representational context*, can organize behavior;
- perception research empirically demonstrates *the process of context coordination* in organizing stimuli;
- Marr's model of vision provides a primitive computational demonstration of *bottom-up categorization*;
- PDP devices misconstrue *the nature of information* as given, rather than perceptually created from stimuli; and finally,
- Edelman's model of *neural map selection* shows how a process memory might work.

At the end, I'll consider particular difficulties readers may have with Rosenfield's statements about learning, goals, and symbols, and provide interpretations that better delineate the opposing points of view.

## 2. What model of memory is Rosenfield arguing against?

Before considering Rosenfield's remarks about localization in more detail, we need to make clearer just what AI and cognitive science have claimed about memory and what specifically Rosenfield rejects. As we have already seen, one difficult idea is that Rosenfield argues that memory and perception constitute one integral process. As we delve further, Rosenfield requires us to alter our views about the nature of concepts, representations, and even information. The process is frustrating because so many related ideas that have been useful in AI research for decades start moving around like ill-defined pieces in a puzzle and merging with each other. The underlying difficult, I believe, is that if Rosenfield is right, we can't say what human memory is like because we have never built anything like it. All we have are bad or misleading metaphors deriving from our existing machines and designed processes (e.g., computer memory, tape recorders, holography).

### *2.1. Not grammars, not stored structures*

To start with a simple idea, it is generally accepted that human memory is associative [2]. But we mustn't assume that because we observe someone associating "CAT" with "DOG" that these words are physically linked by neurological structures in the subject's brain (e.g., that pointers connect places in the brain where these concepts are stored). This is a common way of modeling associative behavior—the brain *appears to behave* as if it were an implementation of a semantic network. Rosenfield is saying that this isn't how the mechanism actually works. A better perspective is that the *process* of saying, seeing, and/or hearing "CAT" is physically related to the *process* of saying, seeing, and/or hearing "DOG". Semantic network models describe how the *perceived products* of these processes (e.g., spoken words) are related. Crucial to Rosenfield's argument, such mechanisms are fundamentally incapable of producing the range of behaviors that neurological processes accomplish.

The semantic network model of memory has been elaborated in the past few decades in what I will call the *grammatical model of cognition*. This approach assumes that knowledge consists of concepts linked in a "memory structure", which is accessed by programs for constructing mental models [42]. Much of knowledge representation research can be viewed in terms of making the grammatical nature of process models more explicit by separate domain models from the inference rules that operate upon them [19].[4] In knowledge engineering, this approach has led to generalization of modeling languages (e.g., causal representations) and inference procedures (e.g, "generic strategies" for diagnosis or design). An implicit claim of knowledge engineering is that human-equivalent intelligence can be produced from grammars.

Although Rosenfield doesn't speak in these terms, his book can be viewed as an argument against the grammatical model. First, programs based on grammars simply follow patterns, they can't break the mold and do something new (essentially the argument of Winograd and Flores [61] about the limitations of representations). We can represent any process by grammars, but if we *replace* the processes of human behavior by grammars we lose flexibility.

A second argument against the grammatical approach—following a quite different tack—is that grammatical models of cognition are based on the localization hypothesis. For example, knowledge engineering (and much of cognitive modeling) assumes that programs in the brain are retrieving and manipulating stored relational networks such as classifications and state-transition networks. Significantly, the argument against stored structures also argues against there being *stored programs* in the brain that are themselves retrieved and interpreted. This is not a distinction between compilation versus interpretation or declarative versus procedural. Rather, what's at stake is the

[4] Knowledge about processes in the real world (e.g., diseases) is represented in the memory structure by a basic set of relations and compositions of them. These relations correspond to links between categories: subtype, cause, part-of, location, time. Problem-solving processes (e.g., how to do diagnosis) can be modeled by rules that refer only to relations, rather than domain terms (e.g., NEOMYCIN's abstract strategy rules [15]). In effect, such rules make explicit the grammar that assembles the domain lexicon into behavioral sequences. For example, the ODYSSEUS student modeling program uses NEOMYCIN's diagnostic strategy as a kind of grammar for parsing a student's sequence of data requests [60]. The relation of discourse rules (e.g., GUIDON's case-method tutoring rules) to content matter (the topics represented by the domain model) is similar; that is, abstracted tutoring rules constitute a grammar for a case-method dialogue.

very idea that the brain *stores* any kind of structures that are selectively accessed and manipulated *as structures*. Once again, memory is not a place where *things* (whether representations or programs) are stored. Rosenfield would say that the compilation model is wrong because there is nothing stored that can be retrieved and compiled into something else.

Putting together the two arguments against grammatical models of cognition, the distinction is not between, say, a program interpreting an equation for a circle and a turtle program for drawing a circle. Both fail to account for novelty and both presuppose storage of structures. Instead, to put it simply, Rosenfield is asking, how do I manage to *trace a new line*, to coordinate what I see with my hand's movement? Again, by the model of memory presented here, every movement is a new coordination, not merely following instructions or executing a program.[5] In summary, grammatical models are rejected because they don't account for the novelty of every behavior and they require stored structures. From the AI research perspective, the novelty argument is of course more important, but the argument against stored structures is valuable because it alone forces rejection of the grammatical approach.

To recap, the conventional view of memory is based on the metaphor of storage. Storage involves putting some*thing* in some*place*. Arguing against localization is not saying that a thing is stored in many places (e.g., multiple copies of "CAT") or that it is distributed over several locations (e.g., "C" and "T" are in very different locations in the brain.) or that it is encoded at another level (e.g., a description of C as a set of curves is stored, not the image of "C" itself). Rather, arguing against localization is more like saying that *concepts are not things*, but processes of perceiving and processes of behaving.

To take a familiar example, human memory is not like a dictionary. Words are not written somewhere in the brain. Words do not have labels or addresses by which they can be "looked up". To be trying to define a word is not to be moving a name around the brain, matching it against other names or indexing the place where it is stored. There are no addresses, no pointers, no labeled networks in the brain. We cannot access and display brain structures and as observers say that they correspond to concepts. Brain structures are not stored away, retrieved, and interpreted as objects. This is the essence of Rosenfield's claim.

### 2.2. *Capacity to compose, cycles of perceiving*

Okay, so what is memory? To begin, we must shift from viewing memory as a place where descriptions are stored to a capacity to do and recompose what we have done before. To use a bad metaphor, contrast a CD as an encoding

[5] Even for so-called rote behavior the neural constructions are new because the external interactional context and the internal on-going context, within which behaviors are coordinated, are never identical on different occasions.

for producing sounds directly with a score representing the music, instruments, and orchestration in some notation. Human memory is more like the capacity to replay what was done before directly; it doesn't require (and indeed *never directly involves*) interpreting a description of what the behavior should look like (e.g., interpreting a score or grammar or following a script).

The CD metaphor is bad because it involves localized encodings of sounds. We don't store a description of how a word sounds or even instructions for generating it. Rather, our neurological structures are biased to reorganize themselves so we can say the word or write it or spell it out again. Human memory is a capability to organize neurological processes into a configuration which relates perceptions to movements similar to how they have been coordinated in the past. If this sounds too vague to implement, that's no surprise: we can't build anything like human memory. We don't know how to describe how it works, just what it *appears* to do.

In contrast with a single utterance, such as saying CAT, conversational speaking, as an ongoing activity, involves intervening use of representations. Conversational speaking involves *cycles* of perceiving, reorganizing neural processes, and behaving. We are "using representations" each time we utter a word or phrase and, reflecting on what it means, make a clarification or elaboration. In describing the process of remembering a word, for example, we must distinguish between the single memory-coordination process of generating a feature (e.g., "it starts with an S") and the overarching cycles of reflecting on feature descriptions—*perceiving them*—and thereby using representations to organize behavior. Structures have to be perceived to be treated as representations.[6]

Elaborating on Rosenfield's argument, we might say that AI and cognitive science have confused the representational manipulation that goes on in the outward behavior of our speaking and writing over time with what goes on *within a single cycle* of categorizing and creating a representation (for example, by uttering a sentence). We must distinguish the mechanism by which we perceive what someone just said and utter a reply from the process over time by which we carry on a conversation.

Donald Schön [52, 62] clarifies these levels of behavior in his analysis of the logic of inquiry, which might be paraphrased as follows:

- *Doing*: Words are used automatically, we are just actively talking (generating representations automatically, but not commenting on them).
- *Adapting*: We are caught short momentarily, but easily continue. We

[6] This means that perception includes interaction with internal constructions, as in visual imagination and silent speech, not just interacting with something outside. Vygotsky [58] considers how these two forms of perception differ, focussing on the development of shortcuts in inner speech.

"glitch" on something unexpected, but respond automatically (automatically commenting on representations).

- *Framing*: What are we talking about? What categorization fits our activity of speaking? We are transforming the conversation (deliberately attempting to generate appropriate representations).
- *History-telling*: We are articulating new theories, relating images to words, describing how we feel, reviewing what has been said so far (reflecting on a sequence of prior representations, composing past perceptions into a new way of seeing).
- *Designing*: We are deliberately guiding the conversation so it becomes an inquiry-project, resolving a problematic situation (defining what representation generation should be about; creating and carrying out an activity involving the above four components to some end; representing what we intend to compose and then managing that composition process).

This analysis makes clearer how representations build on one another. For example, one form of reflection, which I call history-telling (Schön's "reflection on knowing- and reflection-in-action"), involves commenting on a *sequence* of prior representations. Representations play a different *role* in organizing behavior, depending on how they relate to prior behavior. For example, at the base level—within a cycle of perceiving and behaving—*doing* does not involve commenting about representations at all.[7] Crucially, Schön's analysis suggests that representations are constructed compositionally, over time, as the context becomes more complex and subsumes previous observations and commentary. This nesting isn't arbitrary and isn't the same at each level, but has a logical form relating to automaticity, reference, sequence, and functional composition.

The result is a shift in perspective: We view representations as *created in our outward, conscious behavior*—in our imagining, speaking, writing, drawing, not manipulated in a hidden, unperceivable way inside our brains. In its primary manifestation, memory is the capacity for automatically composing processes of perceiving and behaving, including creating representations (doing, adapting). In cycles of such behavior, what James called the "secondary" aspect of remembering, we reflect on (represent) what we have said and done before (framing, history-telling, designing). Thus, memory is fundamentally indistinguishable from coordinated perception and movement—in both its primary and secondary manifestations, relating what we have done before to what we are doing now.

[7] By this analysis, we might say that grammatical models never go beyond "doing" and "adapting". One difficulty in AI research is that what appears to be framing, such as introducing a new topic of a conversation, can be modeled grammatically (e.g., turn-taking rules). Although we have represented these and similar patterns, we have never explained how they develop by interactions between people over time, except as grammatical modifications to grammars, which begs the issue.

### 2.3. Regular behavior without internal representation

One general implication, consistent with Winograd and Flores' [61] analysis, is that human reasoning involves the use of representations (such as this written review), but human behavior is not generated directly from representations. For example, when we speak we are not translating words from an internal description of what we are planning to say. When we do plan what to say, we generate such plans as words or diagrams that we can perceive (including when we talk to ourselves or visualize things). Such plans do not come from other plans directly, but like all speaking and representation creation, they come from our ability to directly sequence, compose, and substitute previous behaviors.

In effect, all speaking involves novel conceptualizations and compositions. There is no internal, grammatical description of sentences that we interpret and apply in some hidden way, just regular ways of behaving (patterns perceived by an observer—abstractions—expressed as representations). Similarly, there is no lexicon of defined words from which our concepts are selected and rotely applied. Indeed, to use a grammatical rule or a word definition, we must recite it first. No representation can be *used* in the sense of being given a meaningful interpretation without being perceived first.

Although this may seem strange to many AI researchers, it is an old idea and has much support in linguistics (Tyler [57]), anthropology (Suchman [56]), and sociology (Mead [37]). Consider for example these remarks by Collingwood from *The Principles of Art* [21]:

> We think that the grammarian, when he takes a discourse and divides in into parts, is finding out the truth about it, and that when he lays down rules for the relations between these parts he is telling us how people's minds work when they speak. This is very far from being the truth. A grammarian is not a kind of scientist studying the actual structure of language; he is a kind of butcher, converting it from organic tissue into marketable and edible joints. Language as it lives and grows no more consists of verbs, nouns, and so forth than animals as they live and grow consist of forehands, gammons, rump-steaks, and other joints. [21, p. 257]

> [A] coagulation of several words into a single whole, quite different from the sum of the words that compose it in their recognized grammatical relations to each other, is called an "idiom". . . . [A]ll the grammarian has done by calling them idioms is to admit that his own grammatical science cannot cope with them, and that people who use them have spoken intelligibly, when according to him, what they say should be meaningless. [21, p. 258]

> Language is an activity; it is expressing oneself, or speaking. But this activity is not what the grammarian analyses. He analyses a product of this activity, "speech" or "discourse" not in the sense of a speaking or a discoursing, but in the sense of something brought into existence by that activity. [21, p. 254]

From this perspective, a blackboard model of the mind, in which discourse plans and sentences are grammatically assembled and posted on many levels of detail before any speaking occurs (e.g., "speakers use the rules to determine how to say what they want to say" (Hovy [27]) is a fantastic reductio ad absurdum account of how speaking actually works. Of course, some AI researchers have realized the implausibility of current models. Minsky [39] suggests that we "put aside most of the old language theories":

> If we're to understand how language works, we must discard the usual view that words *denote*, or *represent*, or *designate*. . . . If we want to understand how language works, we must never forget that our thinking-in-words reveals only a fragment of the mind's activity. [39, p. 196]

The non-representational memory model raises many questions that we thought perhaps cognitive science had resolved, hindering change from old ways of thinking. For example, why does human speech appear to be regular if it is not produced by interpreting grammars? Why do we sense that we are reusing words, rather than forming new concepts? How do we represent and immediately follow rules when we are given explicit instructions (Hadley [26])?

Obviously, there are many stable reconstructions; apparently, the very business of perception is to view the world conservatively (noticing only what is different) in order to adopt previous successful ways of behaving. But although rote recall may be the paradigm of remembering, speaking grammatically and mimicking are hardly marks of high intelligence. Rather it is performances requiring subtle adaptations to apparently new situations, whether on the high trapeze or juggling a financial portfolio, that we view as perceptive and intelligent.

Indeed, even what we take to be a highly stable behavior, such as reciting a phone number, is highly contextual. Phone numbers and log-on passwords are not retrieved, but are speaking or typing or dialing behaviors that occur in the context of other perceptual and motor processes. You can establish this context (a composition of active neurological processes) by sitting in front of keyboard, by visualizing a phone, etc. Rosenfield's main claim is that building a human-like memory requires understanding the integral manner in which perceptions and movement processes are composed, reactivated, sequenced, and coordinated.

To summarize, according to Rosenfield human memory does not consist of addressable, localizable, retrievable structures (stored representations). Rather, memory gives us *the capability to produce structures*, which we call representations, that do have these properties (in our speaking, drawing, writing, gestures, visualizing, etc.). We do not store descriptions of what our behaviors should look like, but rather have the capacity to reorganize our perceptual/motor coordination in ways biased by previous organizations. We don't literally follow a script, though we can create one, perceive it, and organize our behaving accordingly.[8]

### *2.4. Self-organizing on one level, reflective on the next*

The idea of self-organizing, emergent processes is central here, and sharply contrasts with a typical AI architecture for deliberately controlling complex activity. Winograd and Flores [61] describe a committee meeting as an example of a self-organizing process. The members of a committee don't retrieve a description of how to interact in a meeting, which is then executed. Nor is the chairperson determining what individuals say. But the group might have a written agenda, and individuals might speak to themselves, represent what is going on, and plan what to say. These representations are produced so they can be perceived, not manipulated, indexed, retrieved, etc. in some hidden way. First there must be a process of saying or writing something (directly creating a representation), then a perceptual process of commenting on what was represented (reflection).

To say that representations don't directly cause behavior is to claim that hidden interpretation of plans and scripts is not the mechanism that organizes behavior.[9] By analogy, the storage view of memory and representations is like modeling a camera's mechanism by describing the photographs it produces [29]. The dominant AI paradigm is based on the idea that all action follows from descriptions (grammars) that order behavior—as if descriptions of the photographs were inside before any pictures were taken. Cohen [20] had the same problem in designing AARON: How can we build a machine that creates new representations without building in descriptions of them? If the plans, statements, and behaviors in general are produced inside before the actual behavior, from what is this internal description produced? (Could AARON have another agent inside who draws pictures before they are drawn on paper?) To defeat the homunculus fallacy, we must realize that the compositional, modular nature of the mind is in the form of self-organizing processes, not as agents *speaking* to each other, using representations, in the form of

[8] When asked what I think of the CYC project [34], I respond in like vein: When we finally do create an intelligent machine, of which I have no doubt, it will enjoy reading Lenat's encyclopedia.

[9] Cf. Bickhard and Richie's reordering of levels, Section 1.3.

schedulers, message-passing, and agenda. Minsky's [39] society of mind metaphor generally adheres to this restriction.[10]

Progress in AI has been so much based on notions of search and control, we have essentially ignored natural examples of self-organizing processes and what kinds of complex behaviors they can produce. A traffic backup is a good example to start, because it is clearly not deliberately organized. A bottleneck may form where roads converge or narrow. The individual cars are not following a plan for "how to participate in a traffic back-up" or even "how to create today's traffic back-up". The organization that observers see in the lines of cars was not predescribed, but is a structure that emerges through the interaction of many parts. There is no scheduler deciding what car gets to move next. Observers will see patterns in the emergent behavior over time (e.g., as a bottleneck becomes released just beyond the scene of an accident, even hours after the area has been cleared away). But there is no "pattern"—some*thing*—that is being "followed" (interpreted) by the participants.[11]

In general, AI and cognitive science have confused grammatical models, which describe adapted patterns of interaction between individuals over time, with the mechanism that produce momentary individual behavior.[12] That is, we

[10] Minsky says, "Memories are processes that make some of our agents act in much the same ways they did at various times in the past" [39, p. 154]. But on the same page, he says "stores the traces of the past", adopting a storage of substance view, not a process view. In an interesting twist, Minsky suggests there may be two distinguished and specialized high-level agents (the *B* and *A* brains, popularly referred to as the Left–Right brain distinction) that react to the representations each produces, one focusing and articulating distinctions top-down, the other forming images bottom-up (see Minsky [39, p. 59; 43]).

[11] Bartlett uses the example of a game like Rugby football: "Nine-tenths of a swift game is as far as possible from the exploitation of a definite, thought-out plan, hatched beforehand, and carried out exactly as was intended. The members of the team go rapidly into positions which they did not foresee, plan, or even immediately envisage, any more than the bits of a glass in a kaleidoscope think out their relative positions in the patterns which they combine to make" [4, p. 277]. Bartlett goes on to say that if individuals have to think what another player is going to do, the team will be disconnected. In terms I have paraphrased from Schön, there is little place for framing and history-telling during a play. Again, this is not to say that we don't sometimes generate representations to change our behavior, but to underscore that behavior is often possible, indeed required, without them.

[12] Emergent descriptions are a characteristic and necessary aspect of *interlevel* theories. Stable organizations develop over time by interactions between individual parts which themselves can be described mechanistically (e.g., we can describe individual cars in the traffic jam, the goals and plans of the drivers, and how cars locally interact). However, the system as a whole develops patterns that no individual or isolated group could be said to control. As Bartlett explains, the role of individual actions can only be understood in terms of *ongoing trends* of the already organized mass: "We can put our finger upon this, that or the other thing and say: 'This comes from such and such an individual source.' But when we have done all that can be done in this way, there is much left over. It is left, not merely because the phenomena are too complicated, but because any constructive achievement of social organisation depends upon the form and trend of the group before the achievement is effected, as well as upon the efforts of innumerable individuals in the mass" [4, p. 278]. In contrast, *intralevel* descriptions explain behavior exclusively in terms of how individual, connected parts causally affect each other (e.g., how an automobile engine works). *Interlevel* theories relate individual components (e.g., car paths) to systemic patterns or trends

⟶

have described what adaptations occur, but not the local process of adapting. This is why I said that the promise of situated cognition research is to provide better explanations of human learning. The essence of Rosenfield's book is that perceiving, behaving, and learning are one process. We are not retrieving descriptions of what is true or what to do, but constructing (speaking, imaging, moving) behaviors directly from how we have perceived and moved before.

How we talk about reflection in terms of expectations, assumptions, and rationalizations exemplifies our confusion. As Schön [52] says, we engage in "historical revisionism" when we suppose that a "failed expectation" was necessarily represented prior to its articulation. Assumptions are similar. We say, "I did that because I must have assumed. . . ". Such rationalizations are relative to our current context, as we look back as theoreticians and perceive patterns and comment about relations in our behavior, not necessarily something we said before that caused us to act in some way. Winograd and Flores [61] use the term *breakdown* to characterize how representations describing behavior emerge when we seek to explain an interruption in our otherwise automatic flow of behavior ("doing" and "adapting"). Bartlett's experiments illustrate how remembering is a construction or rationalization that finds a way of working around an impasse. It is in framing, according to Schön, that expectations and assumptions are articulated. Such representations are about our activities, but their causal effect is towards the future, as perceptual organizations of behavior [3].

The difficulty in modeling people is that they do use representations and they do represent their own behavior. Every schoolboy knows something about the grammar of his native tongue, and this, we hope, affects his speaking behavior. But we always forget that children speak long before they know what nouns and verbs are. The Platonic view is that categories of speech are inborn ideal forms. Indeed, even modern linguists struggle to account for how the patterns they perceive as observers could possibly be "produced" by subjects without a stored, internal form [10]. They fail to distinguish emergent interactions from preconceived rules. Rather than asking how habits develop as an interaction of perceptions and movements, they continuously wonder how *descriptions of*

observed over time (e.g., bottlenecks) at the level above, not just to physically connected components at the same level (e.g., neighboring cars and roads). Moreover, recalling the fountain example of Section 1.3, emergent structures don't map onto fixed units in the level below (e.g., different cars are caught at the bottleneck at different times). Thermodynamics is a familiar interlevel theory, relating properties and interactions of individual molecules to the volume, pressure, and temperature of gas in a container. The key idea of emergence is that the observed system-wide properties cannot be explained just in terms of the component interactions at the level below—you must refer to the ongoing trends of the system (e.g., the way temperature, a property of a gas volume, affects individual molecules). In the interlevel theory of Bickhard and Richie, representations emerge from the interactions of neural structures in the level below. For Bartlett, the "society of mind" is not just a metaphor; his theory of social organizations closely parallels his theory of neural organizations—indeed, social-neuropsychology is the interlevel theory he strives for.

*how the habits appear to an observer* could be known at birth, encoded, or learned "tacitly". Popular views of instinct and talent as inborn behavior programs are similarly misconceived [5].

In summary, the idea of neurological self-organization is that neurological processes come together in ways that are coherent without a controlling program that assembles neurological structures. The construction is in the perceiving and behaving itself, not a separate process that creates something that is later "run". Again, we don't speak by constructing an internal description of what we are going to say, except of course when we actually do say something to ourselves and reflect on it. A schoolboy doesn't use a grammar representation until someone tells him the rules.

### *2.5. Implications of an alternative model*

Okay, so what? Why should AI researchers care about human memory anyway? We can't build a bird either, but we can get to the moon. Planes don't need feathers. Even if Rosenfield is right, what are the implications for building intelligent machines?

First, Rosenfield argues that previous neurological data has been misinterpreted, distorting our view of memory and knowledge. The idea of modularity of function at the level of reading and writing still holds sway today, influencing both modern neurobiology and AI. Consider for example this remark from a recent book review of *From Neuropsychology to Mental Structure* [22]:

> Some patients whose reading of content words such as *elephant* and *chrysanthemum* is good cannot read even the commonest function words such as *the* or *and*. Examples of such selective deficits are now legion in cognitive neuropsychology. They show that cognition must be profoundly modular. Our semantic systems must have separate subsystems for animate and inanimate concepts; our knowledge of names must involve one subsystem for proper nouns and another for common nouns; and there must be separate lexical systems for content words and function words. These are claims about normal cognition; but they are made on the basis of studies of people with damaged cognitive systems.

Evidence that such interpretations are wrong could suggest new mechanisms to be incorporated in AI programs. Modular separation of data (memory) and programs (the grammatical approach) is surely a virtue for software design, but the very idea of indexing, matching, and assembling structures is the origin of combinatoric search. If the brain has a mechanism for self-organization that avoids the problems of search and matching that are legion in AI programs, we will want to know about it.

Second, as everyone knows, human intelligence grossly exceeds AI program capabilities. This is especially true regarding conceptualization, the creation of new representations. What if we have matters inside-out? What if speaking involves *constantly* creating representations and creating representations that interpret them, as Schön's analysis suggests? What if perception is when learning takes place, as Rosenfield and Edelman claim? By what self-organization architecture are neurological processes subsumed and composed? How is a previous organization reaccomplished more easily (the practice effect)? AI researchers know that adaptation is important. Learning is roundly acknowledged as our key unsolved problem. We might look more clearly at psychological data (e.g., [9, 31]) and realize what aspects of human flexibility are not captured by grammatical models, and then find alternative models for capturing the underlying processes (for example, see the interlevel architecture theories of Iran-Nejad [28] and Bickhard and Richie [7], mentioned briefly in Section 1.4).

Third, when we study knowledge bases such as NEOMYCIN's, we find that Collingwood's claims (Section 2.3) are not so foreign. In effect, situation-action rules model what experts most directly know—*how to behave*. But when we study and decompose these rules into conceptual and procedural abstractions (such as occurred in the transition from MYCIN to NEOMYCIN), we are stating a nodel that goes well beyond what experts state without our help [15]. If human memory is not a place where representations are stored, this new understanding could have a dramatic impact on how we view the knowledge acquisition bottleneck, as well as explanation and teaching using knowledge bases [18].

With this background, I now turn to the main sections of Rosenfield's book.

## 3. Classical memory research

Rosenfield's reinterpretations of seminal experiments and case studies of neuropsychology illustrate alternative—non-storage and non-representational—explanations for what we term recall, recognition, and skilled performance. His main claim is that failures to behave appropriately reveal much more about human memory than "failure to retrieve" or "failure to execute a plan". We will consider here some of the key researchers in Rosenfield's review.

### *3.1. Paul Broca*

Paul Broca is perhaps most remembered for localizing speech capacity to an area in the brain. However, in 1861, Paul Broca argued "not (for) a memory of words, but a memory for the movements necessary for articulating words" (p. 20). Because of this emphasis on coordinating movements, Rosenfield views

Broca's work in a positive light. Unfortunately, Carl Wernicke in 1874 and Ludwig Lichtheim in 1885 took Broca's localization of speech to an extreme by emphasizing the idea of separate centers where visual, auditory, and motor "images" are recorded. In particular, they argued for stored records of individual words. Yet other studies by Armand Trousseau in 1865 had already indicated the importance of context. For example, a patient could not repeat a word like "student", but could repeat "all the students". This suggested that "student" was not represented in a single place as a single word in the brain.

### 3.2. *Frederic Bateman*

In 1882, C. Giraudeau described a case of *word deafness*, by which a patient could hear spoken words, but not recognize what they mean. In 1890, Frederic Bateman interpreted this problem in terms of "complete loss" of function (and hence dysfunction of a localized part of the brain). Rosenfield argues that the same data can be interpreted in terms of difficulty in establishing a context, that is, an inability to coordinate, to get a coherent process moving in the current situation, relative to what was done before. For example, rather than saying a patient can't understand speech, the patient apparently "has great difficulty establishing a context in which she can understand the questions being asked" (p. 29). For example, when asked her occupation, she describes her medications, as if she anticipates the logic of the inquiry, but is unable to coordinate her memory of the process with what is currently happening. She apparently has some sense of a familiar setting, but can't dynamically correlate her past experiences with what she is currently perceiving; she cannot construct a coherent new composition that is analogous to what she has done before. Given that she can speak and does after much repetition answer questions appropriately, explanations of her deficiency in terms of individual words would appear to be inadequate. Nevertheless, this is what Bateman concluded.

### 3.3. *Jules Dejerine*

By the time of Jules Dejerine's work in 1891, the localization of function was well accepted. Dejerine explained reading failures as disconnections between visual (right hemisphere) and language (left hemisphere) faculties. One patient could "recognize" only single letters or two-digit numbers (but not words or more complex numbers); he could recognize a signature but not the individual letters in it. The patient demonstrated an intriguing ability to state letters when he made the corresponding writing gestures. Dejerine concluded that "motor activity (writing) can organize stimuli, making recognition possible" (p. 58), which Rosenfield cites as a theoretical advance.

However, Dejerine's basic theory was that the patient couldn't write because he couldn't read, and this implied a *word blindness*, a retrieval or disconnect problem. According to Rosenfield, Dejerine makes a false distinction between drawings and symbols, as if the recognized forms were treated as pictures being matched in the brain. The very idea of word blindness wrongly suggests that the patient receives words as stimuli, and as such they are unrecognized wholes. Instead, the failure to read words reveals an inability to construct "an overall organization—in which identical stimuli, letters, are constantly changing in significance" (p. 49). This is not a breakdown of a specific linguistic function, but a general inability to organize stimuli, a coordination or composition problem. To summarize:

> If recognition depends on being able to organize similar stimuli in a variety of different ways, memory, too, must in some sense be based on this organizing capacity. When we recognize a face, we are organizing stimuli in ways that are similar (but not necessarily identical, since the person might have aged) to how we have organized related stimuli in the past. It is the similarity of organization that relates past and present. (p. 50)

### *3.4. Norman Geschwind*

More recently, in 1965 the late Norman Geschwind refined the disconnection hypothesis by explaining capacities such as reading as composites of independent brain operations. Information from different brain centers is available for correlation, explaining why one sensory association may fail when another succeeds. For example "a man can see an object without recognizing it, and yet he can touch the same object and have no difficulty naming it" (p. 56). Such cross-modal associations and associations of associations are generally compatible with Edelman's models. But Rosenfield is always skeptical of claims that parts of the brain are working independently of each other and produce results that must be later assembled. Geschwind's cross-modal view was substantiated by Charcot's earlier experiment that showed one could read by tracing letters, suggesting a graphic-motor center. But Rosenfield emphasizes that Dejerine's later interpretation of his patient, Oscar, didn't require such a center to be postulated. Dejerine pointed out that one can write with different hands, which clearly aren't controlled by a single "writing center" (given that localization of left-right muscular control is accepted).

### *3.5. Appraisal and discussion of the historical view*

With all the many historical cases and multiple interpretations, Rosenfield's survey would benefit from a table or time-line summary. The exposition is confusing at times because many of the researchers are inconsistent from Rosenfield's perspective. For example, Dejerine attacked the idea of a writing center, but still held to the idea of fixed visual images of words stored in different places.

We come away with a wide variety of interpretations of fascinating clinical cases. Although it's not the best scientific history, it is sufficient to make Rosenfield's point that there are alternate models of performance failure, and the fixed-trace/localization hypothesis is not only unnecessary, but obviously simplistic. Other explanations for dysfunctions are possible:

- a previous organization between sensations and movements can't be reaccomplished (visible objects can't be named);
- a process can still be accomplished at a high level, but it can't be integrated with ongoing processes in the present (the patient carries out a medical interview, but out of synch with her inquirer);
- a past neural organization was never retained at all.

To make this more concrete, contrast these views with a typical AI interpretation of Loftus' studies of confabulation on the witness stand [35]. According to the stored memory view, the witness has a permanent record of what she actually observed at the scene of the crime. If this record isn't manifested in a correct recall performance, it is because her internal description is transformed or distorted by the *pragmatics* of the courtroom situation. She actually knows what she saw, but her current biases lead her to produce a new story. Or perhaps resource problems (not having enough time to remember) prevent her from retrieving the truth, so she fill in the details. Hence, the AI model is RETRIEVE + FIX, rather than CONSTRUCT + FIX.

A good example of the RETRIEVE + FIX model is Repair Theory (Brown and VanLehn [12]), which postulates that subtraction bugs arise from impasses caused by omissions from an otherwise correct procedure. A student knows the ideal form of the subtraction procedure, but there are gaps in what he can retrieve, so he makes those parts up. One problem with this grammatical model (cf. Section 2.1) is that bugs change over time. Behavior isn't identical every time, as a rote view of memory would imply, so the notion of selective forgetting and "bug migration" must be introduced. Here we have the equivalent of epicycles in cognitive modeling.[13]

[13] In general, "pragmatics" is invoked to explain why behavior doesn't adhere to grammatical descriptions, or more generally, "why and how is it that we say the same thing in different ways to different people, or even to the same person in different circumstances" [27]. The general claim is that pragmatics take into account context, in the form of the agent's relation to his environment (e.g., conversational atmosphere, interpersonal relationships, goals to influence behavior). But if oral speech is fragmented because of the pressures of real-time interactions, why is written speech generated in pieces, scratched out, restated, reordered, and reconceptualized? Text isn't created whole and spit out in a single stream. We reflect, speak, reflect, revise. The idealized "text generation" of natural language programs never occurs in human experience: Speech is fragmented and written prose is revised. Why do we produce "inadequate" prose? Why can't we say it right the first time? Here is the essence of what representations do for us: *We don't know what we want to say until we say it.*

The primary difference between the retrieval and construction models is not in the cycles of reflection and commentary on the evolving story, which is essentially the level that most AI programs model, but where each utterance comes from. Rather than *retrieving* descriptions of what happened and translating them into words (or retrieving procedures to be executed), the constructivist view is that *speaking is conceiving*:

> When Dejerine stressed that writing could be carried out in a number of different ways and therefore is not localized, he, too, was suggesting that memory is procedure, but without seeing this fact as true of all motor acts, including speech. (p. 62)

In contrast, Bartlett [4] described an exclusively constructivist model based on his data of story recall experiments. Bartlett suggests that perception and speaking are a single process (the categories called *doing* and *adapting* in the discussion of Schön, Section 2.2). Impasses occur when automatic behavior is unable to continue forming a coherent composition (a story). Bartlett describes a process of resolving the gap by which a detail, usually an image, is perceived, an act that forms a new composition (corresponding to *framing* and *history-telling*, Section 2.2). Comments about the meaning of these perceptions—now treated as representations, that is meaningful forms—constitute categorizations that orient subsequent behavior.[14] Note that Bartlett's model of memory requires a different view of perception, concepts, and representation. The bundling of these ideas—which now become ill-defined and moving pieces of a puzzle—is what makes shifting from the memory-as-structures perspective so difficult.

Relating this back to the clinical studies, Rosenfield is attacking the view that drawing a letter is copying an internal description of it, or that speaking is translating an internal description of what we intend to say. By this grammatical view, there is always an internal representation (plan or image) that precedes and controls movements. Rosenfield takes a procedural (process) view: "The procedures of writing themselves create the 'images' we set on paper" (p. 61). Furthermore, this is not merely a form of knowledge that is "compiled" (because there are no structures to retrieve and compile), but must be the basis of all behavior, including declaring facts about the world. Nor are these procedures stored programs that are retrieved and applied to current information: They are *processes constructed on the spot* that correlate stimuli and movements. To develop the view that memory is inherently a capability to coordinate and construct such processes, Rosenfield turns to Freud, Marr, and Edelman.

[14] Again, "categorization" suggests a process; "category" suggests a static description. The emphasis is on recurrent processes of organizing behavior, not stored descriptions of the world or how behavior appears.

## 4. Freud

Rather than making emotions something secondary to an ideal memory, Freud viewed them as controlling behavior. Emotions play an integral role in setting up the context by which ongoing behavior is composed:

> Crucial to the Freudian view is the idea that emotions structure recollections and perceptions. (p. 6)
>
> Recollections without affect are not recollections. Emotions... establish a memory's relative importance in a sequence of events much as a sense of time and order is essential for a memory to be considered a memory, and not a thought or a vision at some particular instant, unrelated to past events. (p. 72)

Where does this leave memory theories like MOPS, which do not include emotion? Relative to the process of human memory, most cognitive science theories appear to be more akin to database maintenance, rather than human psychology. To treat emotion as a veneer—a coloration of something fundamentally concerned with storage and retrieval—is to adopt the Platonic view again of ideal forms that emotion only distorts, rather than processes that an emotional orientation creates and reconstructs. "Events that become emotionally charged are thereby categorized and 'understood'" (p. 73). Notice that the grammatical approach only allows emotions to be labeled and stored, just like concepts. It is probably because this is so intuitively disagreeable that emotions have been routinely omitted from AI models. The rationalist view is that emotions can only distort logical thought; real thinking only occurs when we don't allow emotions to get in the way.

Freud's work suggests that emotions are perceptions that make a kind of commentary on other perceptions. For example, Bartlett claimed that his subjects experienced an emotional attitude about their on-going story-telling, which was correlated with resolution of an impasse. The emotional attitude signified a "coming to terms" with a subject's overall state. According to Freud, in our ongoing sense-making an emotional experience leads a perception to be attended to, named, and thereby remembered. Put simply, emotions play a pivotal part in explaining how habits of seeing and doing are formed. Emotions are important in understanding cognition because they are evidence that non-representational "memories" can structure behavior.

But Freud still held to the idea of memory as a permanent record. Arguing against Freud's interpretation of dreams, Rosenfield says,

> [Dreams'] lack of sense is a lack of context, not disguise and displacement. The mechanism of condensation is an illusion created by a [later] interpretation in which one seeks a context that can give the image meaning and coherence. (pp. 75–76)

Thus, relative to the interpretation, the dream is condensed, but it was not *produced* from this interpretation, like an abridged or condensed book. Dreams, according to this view, are ambiguous fragmentary constructions, "because there are no constraints on the organization of these fragments" (p. 75). Whatever conscious control we exert in forming an ongoing, overarching composition from our perceptions while awake (what Rosenfield repeatedly refers to as coordination of sequences of perceptions and movements) is evidently not active while we are dreaming; in some sense, we are not paying attention to what we are doing. Interpreting a dream is supplying a context that resolves the ambiguities and unifies the fragments into a whole (p. 76).

Here Rosenfield summarizes the alternative to a fixed-trace memory:

> There are no specific recollections in our brains; there are only the means for organizing past impressions. . . . Memories are not fixed but are *constantly evolving generalizations*—recreations—of the past, which give us a sense of continuity, a sense of being, with a past, a present, and a future. They are not discrete units that are linked up over time but a dynamically evolving system. (p. 76, emphasis added)

Realizing the centrality of sense-making, Freud postulated the unconscious as an agent responsible for maintaining "the dynamics of the categorizations and recategorizations that give our mental life the sense of a whole. . ." (p. 77). But Rosenfield argues that specific unconscious memories, supposingly what the dream is dredging up, "would not account for our sense of continuity; continuity is a consequence of our ability to view things in larger relations given the present."

Indeed, Freud pursued this point of view early in his work, in *On Aphasia*, illustrated by his somewhat startling remark:

> "Perception" and "association" are terms by which we describe different aspects of the same process. But we know that the phenomena to which these terms refer are abstractions from a unitary and indivisible process. . . . Both arise from the same place and are nowhere static. [24, p. 57]

Cognitive science theories have not been built on this insight and the clinical studies from which it came. Similarly, structure-based models of memory ignore Bartlett's insistence that schemas are not fixed traces that are stored and retrieved, but always constructed during the remembering process. Modern theories view perception as a *peripheral* process that parses the world into discrete structures (called "symbols"), upon which association operates, manipulating and storing them away as other static structures, which remain unchanged until they are used again.

Although Freud believed there to be permanent traces, his mechanisms emphasize that memories are new creations. For example, the obsessional neurotic's "redoing" of unpleasant experiences, attempting to make them into pleasant experiences, is doomed to failure because they are driven by unpleasant feelings. Rosenfield points out that these attempts are often realized as physical acts in "a very real attempt to redo, to create, a memory" by making some ritual movement. Ritual *activity* of this sort underscores the connection between memory, perception, and movement. Rosenfield generalizes this:

> In fact, we are all "redoing" the past, and an act of repetition must be understood not as an act symbolizing a specific past event but rather as a whole history of attempts at recapturing the past, a history that is being put into a specific context at a given moment when the repetition is occurring. . . . Just as it would be misleading to say that a pianist's rendition of a sonata is a recollection of an earlier performance. Every performance is unique, though every performance does have a history; but the significance of that history depends on the present context. . . . [T]here is no recollection without context. And since context must, of necessity, constantly change, there can never be a fixed, or absolute memory. Memory without the present cannot exist. (p. 80)

"Context" here, like information and memory itself, becomes slippery and dialectic in nature. It simultaneously refers to "what you are currently doing", "what you perceive to be happening", and "the on-going internal construction of neural processes". Context is neither inside nor outside, past nor present. It constantly changes. Remembering does not exist apart from behavior in some context. Every performance coordinates what the person is currently doing with what has been done in the past. "Memory" is not something fixed or absolute; it is only manifested in the context of a performance and each manifestation is different by virtue of being adapted (cf. Bartlett's remarks about tennis. Section 1.4).

The nature and importance of context is supported by further reinterpretations of Freud and an appraisal of the work of A.A. Low, and A.R. Luria. Misinterpretations of Low and Luria's results on memory for nouns, verbs, and function words influenced modern ideas about modularity (as cited in Section 2.5). In fact, Low concluded that it is not the grammatical category that matters so much as the difficulty of establishing context.

> Words like *at* and *as* have a range of meanings (at home, at ease, at your service) and acquire a specific sense only in a given context. On the other hand, words like *beyond* and *above* have definite meanings of their own, regardless of context. (p. 86)

Of course, we can only say what "beyond" means by supplying a context for interpreting it; but the interpretations appear to be similar in different contexts. Context-generality could account for why patients can read "content words" like *elephant*, but not "the commonest function words such as *the*" [22]. Storage by category need not be invoked as an explanation. Rosenfield describes other, more contemporary work that ignores the effect of context (the lexical dictionary approach of John C. Marshall and Freda Newcombe) or still adopts a filing system model (the verb concept hierarchy of Elizabeth Warrington).

## 5. Perception research

As stated in the introduction, Rosenfield attacks the view that "there is no perception without prior 'learning'. . ." (p. 7).

> Nobody pretended to understand the mechanisms that created the fixed images. That is a physiological question; its resolution would tell us little or nothing about the nature of memory. (p. 15)
>
> Furthermore, a hidden and unquestioned assumption of the localizationist view is that there is some specific information in the environment that can become the fixed memory images. But if recognition depends on context, it is the *brain* that must organize *stimuli* into coherent pieces of information. . . . [F]unctional specializations, suggested by the study of clinical material must be illusory, for what is implied is not that the brain creates our perceptions out of ambiguous stimuli but that it *sorts* neatly packaged information coming from the environment. (p. 63)

From Rosenfield's point of view, you haven't explained memory at all unless you begin with perception of stimuli. Information is not given but created ("as categories, organizations, and orderings of stimuli" (p. 66)) (cf. [45]). This is the inherent flaw of cognitive science and most of connectionist research today. The world does not present itself as interpretable symbols. To predigest the world for your program by prelabeling things and events is to bypass the essential problem that memory must address.

The process of perception must begin at the level of stimuli; it must create the representations that reasoning operates upon (recall the *within* versus *sequence of cycles* distinction of Sections 2.2 and 2.4). This creation is inherently a process of categorizing in every situation, because perception is inherently contextual and the context—an ongoing, internal composition by which perceptions and movements are organized—is always new.

The book gives fascinating examples of experiments that establish that "sounds are categorized and therefore perceived differently depending on the presence or absence of other sounds". For example, there is a "trade-off between the length of the *sh* sound and the duration of the silence [between the words of "say shop"] in determining whether *sh* or *ch* is heard" (p. 106). In fact, "lengthening the silence *between* words can also alter the *preceding* word" (p. 107). For example, "if the cue for the *sh* in 'ship' is relatively long, increases in the duration of silence between the words ["gray ship"] cause the perception to change, not to 'gray chip' but to 'great ship'." Hence, phonemes are not *given* but constructed within an ongoing context of overlapping cues. "What brain mechanism is responsible for our perceptions of an /a/, if what we perceive also depends on what came before and after the /a/?" (p.110) In no sense does an /a/ exist somewhere in isolation in the brain.

The basic claim is that "the categorizations created by our brains are abstract and cannot be accounted for as combinations of 'elementary stimuli'." There are no innate or learned primitives like /a/ to be found in the brain; that is, there are no primitive stimuli *descriptions* in the brain that can be combined. There are just patterns of brain activity that correspond to *organizations of stimuli*. Our perception depends on past categorizations, not on some absolute, inherent features of stimuli (such as the frequencies of sounds) that are matched against inputs (p. 112).

But the memory-as-structures approach holds that:

> . . . acquired knowledge is stored as fixed images in specific centers, just as the nineteenth-century neurologists believed. . . . The world is knowable according to this view, only if it is already known: the recognition of a shape is possible only if there is a fixed image of that shape already stored in the brain. (p. 112)
>
> Seeing, they [AI researchers] argued, requires first knowing what one is looking for. (p. 115)

David Marr challenged this view. Rosenfield reexamines Marr's approach in order to contrast it with other AI research and illustrate how perception can be modeled computationally as a constructive, bottom-up process.

## 6. Marr

Elizabeth Warrington's work, which was briefly mentioned in Section 4, "suggested to Marr that the brain stores information about the use and function of objects separately from information about their shape, and that our visual system permits us to recognize objects even though we cannot name them or describe their function" (p. 117). This separation of function continues the "localizationist orthodoxy", but the idea of recognizing shapes without named, stored images is a dramatic departure. Marr's approach is to view the world from the organism's perspective, in terms of what stimuli it is given and what it must be able to accomplish for its goals. He showed how shape can be derived from intensity correlations, independent of a memory system describing what shapes might be present. But Rosenfield points out that Marr's programs "were carefully designed, rather than learned, suggesting that Marr had not fully solved the problem" (p. 122). Although Marr pointed a way out of requiring a fixed memory of images upon which perception operates, he didn't fully develop these ideas and indeed used them "to justify arguments for functional specialization, overlooking [the] radical implications".

In summary: "In its failure to free itself from fixed symbols, the computational approach ultimately stifles what could have been a new view of memory as procedure" (p. 128) (cf. Bickhard and Richie's "encodings hanging in mid air", Section 1.3). Even though Marr's programs could recognize a shape bottom-up, "the naming of the shape [e.g., as a cube] still relies on access to a fixed memory" (p. 126). Marr's view is that the recognition procedure is fixed at perception time (given) and information is objective (given, perceived by everyone); thus recognition is still a *match* between internal descriptions and external stimuli. More flexibility is required: Perception and learning are inseparable; one cannot be input to the other. Perceiving is learning. But how is experience (what procedures were used previously) part of perception? The key process to be explained is how context is established by building on what the organism perceived in the past, without requiring this experience to be stored as symbolically interpretable, fixed structures (including programs). Connectionist research attempts to address this issue.

## 7. PDP devices

Rosenfield rejects parallel-distributed processing (PDP) devices [48] as a solution to the memory problem because they hold to the idea of a fixed memory, manifested by static storage of weights: Every item is represented by a specific pattern of activity (p. 147). A PDP machine constitutes a clever hash-coding scheme, based on the same idea of conventional computation, in which memory is a place for storing things. The memory's distributed nature does not change this fundamental characteristic.

It might be objected that the contribution of connectionism is not at the level of how categories are stored, but of forming new categories. But according to Rosenfield, the generalization capability of PDP devices is "prefabricated" by the programmer's encoding of inputs. The machine's capability to associate a color with an unknown flower originates in the programmer's encoding of the flower in terms of codes similar to those already learned (e.g., supplying tokens for size and shape), not in the nature of flowers or colors as encountered in the world (p. 148).

In contrast, categorization depends on not just finding common subfeatures of stimuli and noting their contextual relations (p. 148), but on combining stimuli in a useful way (that is, related to motor activities, sense-making and emotions). Furthermore, what constitutes information for an organism cannot be given by a teacher, but must arise from the organism's own organizing processes in interaction with its environment:

> The generalizations in PDP devices are nothing more than overlapping patterns in *predetermined* codes. Real generalization creates *new* categories of information. . . from the organism's point of view, the consequence of unforseen elements in the environment. (p. 149)

For example, a PDP machine, unless preprogrammed to do so, cannot recognize a smudged letter as a symbol. A PDP researcher might claim that the preloading of the net in a teaching phase is just a way to get started, to show how perception occurs based on experience. But this begs several questions: Why does the organism attend to particular stimuli at all? How does past experience influence the perceptual process itself? "How do we create new ways of viewing the present, new *kinds* of generalizations?" (p.. 152) "How [do] patterns of activity acquire significance in a particular context?" (p. 153) For example, how do children acquire the idea of past tense? (p. 155) Edelman's work addresses these issues by suggesting how categorization is a perceptual process.

## 8. Edelman

One of the key ideas in Edelman's developmental approach is that "the nervous system can only approximate what it has already produced" (p. 220). In dismissing the idea of an ideal recording device, we must also change our idea of past and present: There is no time stamp in memory, only correlations and sequences of states (organizations) (p. 162). "What [the patient] lost was not time but the way events and objects were related. . . . There are no calendars in the brain". Rather, a sense of time depends on the ability to construct a context that composes the sense of what is happening now with previous behaviors. That is, placing the past in perspective requires a kind of coordination between past processes of behaving and current stimuli:

> Memories, then, are the procedures that are responsible for the organization of perceptions. They are themselves *generalizations* of previous experiences, ways of organizing sensory stimuli that permit them to be related to past experience. (p. 62)

Memory is a correlation and coordination process. In reinterpreting the clinical studies, Rosenfield finds that:

> [I]t is not that memories have been lost, but that the ability to establish correlations has been destroyed. Those utterances that require little or no such ability remain, giving the illusion that a few specific memories [such as recognizing individual letters or writing one's signature] have been spared destruction while all other words have been lost. (p. 71)

So how does this correlation process work, according to Edelman?

> [N]euronal groups are organized into sheets, called *maps*, and the interactions among the numerous maps—and the fact that all maps are connected to a motor output and to the initial sensory input—categorize information. *The past is restructured in terms of the present.* Perception and recognition, then, are part of the same unitary process. (p. 9, emphasis added)

Behavior isn't determined by single maps (between sensory and effector systems) but in the *relation* of maps to one another at a given time. Again, the cause of behavior is not localizable to specific structures that are permanently associated with that behavior. In a lucid description of Edelman's theory, Rosenfield raises the provocative possibility that neural transmissions are establishing boundaries between neural groups (structural sheets). They are not *communicating* anything via the electrical pulses (i.e., transmitting symbolically interpreted information), but rather are establishing demarcations that constrain further additions and modifications of boundaries. Thus, an organization of neural processes would be composed of a combination of (multi-dimensional?) boundaries. One can further speculatively imagine hierarchical mappings built up as sheets reorganize themselves within the spaces created by currently active sheets. Some key ideas are *reentrant maps* (those that feed back on themselves, allowing for their stimulation in the presence of later different, but similar inputs), *maps of maps* (secondary structure of neural groups), and *multisensory intersections* (correlating maps from different sensory systems for coordinating movements).

Edelman's model of the brain starts with an initial population of maps between sensors and effectors. These maps are then selectively reinforced by use, in direct analogy to the selection of antibodies in the immunosystem. The model has been developed in a series of computer programs; an example in an appendix illustrates how cross-correlation of mappings leads to generalization. Further details can be found in Smoliar's [54] review of Edelman's book.

Applying neural mapping ideas to situated automata [36], we might *model* neural selection in terms of finite-state automata that are selected (as opposed to assembled from abstract descriptions or schemas) by use. Layers of inhibitory and excitatory connections might emerge (each layer produced by a boundary?). Similarly, we might work backwards from models such as MOPS to see what kind of composition of maps corresponds to the observed contextual aspects of reminding. For example, could the failure-driven nature of memory observed by Schank be generalized in terms of perceptual impasses?[15] In general, the book opens the door to reconceptualizing the patterns represented in cognitive science models, so they can be integrated better with the psychology of memory and perception (cf., [41]).

Rosenfield concludes that "The brain is biological structure. Only in terms of biological principles will be able to understand it" (p. 10). This remark is too strong and probably wrong. Surely functional descriptions of intelligent behavior (for example, existing cognitive science descriptions of complex problem-solving and discourse) and studies of the environment emphasizing social structures (e.g., [32, 33]) will be useful for characterizing what the brain accomplishes, and hence help us to comprehend what the neurons are doing. Indeed, the notion of categorization so central to Edelman's theories is arguably in the psychological domain, not biology alone. Rosenfield might have better said, "Only by *incorporating* biological principles will we be able to understand the brain". A social view of learning without neuroscience is like attempting to explain family resemblance without molecular genetics. But the chemistry of amino acids itself could neither predict nor explain population dynamics or punctuated equilibria [13]. A balanced perspective will give proper attention to each level of analysis (neural/genetic, representational/phenotype, social/environment) and describe their interplay.

## 9. Criticisms of Rosenfield

To make Rosenfield's arguments about localization understandable for an AI audience, I have focussed on the distinction between memory as stored structures and self-organizing processes, which is what I believe Rosenfield means when he says memory is procedural. Similar problems arise when we consider Rosenfield's discussion of learning variability, goals, and symbols.

[15] By hypothesis, whatever is perceived is learned and hence "remembered" [49]. But most perceptions, by the very process of construction from past organizations, are similar to past experience (i.e., made analogous by subsumption and composition of maps). As James said (Section 1.4), experience is full of recurrence that doesn't provoke a secondary experience of remembering. To be distinguished, an experience must resist categorization. We remember our failures because that is when coherence, our story-telling accomplishment, required deliberate framing and history-telling to guide the categorization process. That is, as Bartlett, Winograd, and Schön describe, it is at an impasse – an inability to act and adapt automatically – that we use (generate and perceive) representations in order to behave. We later reconstruct (remember) the representations that eased the original impasse. It remains to explain how horses and pigs get by without this.

### 9.1. *Variability from emergent behaviors*

Rosenfield often makes strong statements intended to attack the grammatical approach of AI programs. But he is apparently unfamiliar with machine learning research and therefore fails to make clear the limitations of current programs. For example, he argues that AI programs are inflexible, identifying it with "genetic determinism", the idea that genes constitute *instructions* for assembling the body. He claims that this "strains credibility because it makes it difficult to account for the enormous variability of thought and action" (p. 171). Those familiar with the capabilities of AI programs to plan and assemble new action sequences may be tempted to respond, "but we have programs with such flexibility". Rosenfield says, "Accurate memory traces would hardly help us survive in an ever-changing world" (p. 79). At first glance, an AI researcher might conclude that Rosenfield doesn't understand how schemas can be composed and adapted to new circumstances.

Rosenfield's claims about variability are important, but they don't come to grips with the essential difference between stored-program and self-organizing systems, which he should have emphasized (recall the examples given in Section 2.4). In a stored-program system, a controlling process directs how structures and behaviors are organized according to template (schema, script, grammar) descriptions. In a self-organizing mechanism, global patterns of phenotypic structure (the physical appearance of the organism) and behavioral routines develop over time from local interactions between internal and environmental processes. Bateson [6] describes how stable organizations (physical and behavioral) can emerge through the interaction of two stochastic processes, one with a digital randomizer (genes, neurons), the other continuous (the environment). Such mechanisms are explored by "artificial life" research (e.g., [55]).

### 9.2. *Goals as ongoing, constructed processes*

Similarly, Rosenfield's claim that goals determine the kinds of information that the brain is capable of deriving from environmental cues (p. 121) at first glance appears consistent with AI views. Rosenfield doesn't realize the complex ways in which goals, focus of attention, and behavior interact in today's complex programs. He thinks he is laying out specifications that no program could approach, when in fact on the surface these are the very concerns of everyday AI research.

Rosenfield fails to emphasize that a goal, in his model of memory, isn't tied to a description of something to be matched in the world. A goal should be thought of as an active, organizing process (cf. Bickhard and Richie's, "interactive control structures" [7]). According to the composition idea, a goal is a perceptual categorization that orients the construction of a new perception and hence ongoing movements. (For example, see Schön's [52] description of how ways of talking subsume ways of seeing in the invention of a synthetic paint brush). A goal is not a label or a program, or even something different from a conceptual category, except for its status as active and dynamically orienting behavior. Observers (either theoreticians or subjects themselves) ascribe goals to activities in their framing, history-telling, and reflections on design. Talk about goals, like other representations, must be perceived to effect behavior (recall the discussion of grammars and plans in Section 2; see also [1, 56]).

Rosenfield goes on to say that physical attributes of the world are "a consequence not of any programs in the brain, but of our experience in the world" (p. 135). Readers familiar with machine learning programs might object that a program can change how the world is described by virtue of experience. What Rosenfield apparently means to say is that a program that is fixed (when perception begins), rather than constructed on the spot as part of an ongoing sense-making process, is insufficient. In particular, if a program is always required for perception, the stored-program approach begs the issue of where the first programs came from. Again, this is not the distinction between "declarative" and "procedural" knowledge as discussed in the AI literature. Both declarative and procedural views suppose that perception is carried out by structures that are fixed before the perceptual activity begins and only changed after the perceptual activity is complete. But if there is no protected place where this program could be stored (if memory is not a permanent record and functionality is not localized), procedures must be changed during the perceptual process itself.

Again, what is at stake here is the mechanism: stored-program versus self-organizing system. By analogy, the stored-program idea says that genes are predetermined, internal descriptions of how the organism will appear (declarative view) or they are predetermined, internal instructions for how to assemble structures (procedural view). Similarly, the stored-program idea says that behavioral routines are generated either by interpreting descriptions (e.g. scripts) or by running a stored, assembly program. The self-organizing view requires a kind of machine we haven't yet built (indeed, there is reason to believe that we must first change our idea of what a mechanism can be). The self-organizing view is highly relational, dynamic, and interactional. Information isn't given, it is created. It is the relations between stimuli and differences that are detected, not individual, objective things (something Bateson [5] constantly emphasized). What is "stored" is what neural maps have been active in the past in relation to other active maps.

### 9.3. *Symbols in the brain*

Rosenfield nicely summarizes the lesson from Edelman: "We perceive the world without labels, and we can label it only when we have decided how its features should be organized." However, he may go too far when he says, "There are no symbols in the brain; there are patterns of activity, fragments, which acquire different meanings in different contexts" (p. 166). This is apparently a contradiction. For a pattern of activity (a perception) to acquire meaning is for it to be symbolic, to be treated as a representation. Hence, when I interpret something I have just imagined or perceived (e.g., silent speech)—an activity that clearly occurs in my brain—I am treating patterns of activity within my brain as symbols. What Rosenfield means to say is that representing (e.g., saying something), producing forms that can be perceived, occurs at a higher level, *in sequences of behavior*. In the case of silent speech, for example, we are not literally producing *sounds*, which is perhaps Rosenfield's point. But our experience is the same as if someone spoke (sound hallucinations). We "hear" the sounds, then interpret what they mean. Representations are thus created and interpreted in *cycles of perceiving* (recall the levels of reference in Schön's analysis of framing, history-telling, designing). Representations are not manipulated, stored, indexed, etc. *within* each perceptual act. Symbol structures—meaningfully interpretable forms—must be produced so we can perceive them, and this occurs in our writing and speech, as well as privately in the brain.

## 10. Common objections

In the process of preparing this review, I have been repeatedly asked a number of questions, which are summarized here.

**Question 1.** *You refer to AI of the past. Recent work in connectionism and parallel processing is addressing these issues.* There is indeed no reason to draw a boundary between AI, situated cognition, connectionism, etc., causing researchers to feel isolated or obsolete. It is important to build on the insights that arise from different fields with different motivations. For example, Ullman's architecture for bottom-up vision is exploited and extended in Chapman's [14] model of situated cognition. However, we must be clear that talk of "connectionist symbol mapping", "language of thought", and "parallel, distributed problem-solving" generally adopts the metaphor of memory as stored structures. These are not alternatives to the grammatical approach.

**Question 2.** *It's not new.* In part that's my point; the information-processing view of the mind has failed to take into account psychological and social theories developed over the past 70 years. Related claims that "we tried that and failed" (e.g., pattern recognition) miss the point that situated cognition research builds on cognitive science; it's not just rehashing old ideas. We want to integrate cognitive science, knowledge-level models (e.g., prototypes, novice/expert differences, misconceptions, strategies) with the psychology of perception, psychiatry, and the social sciences.

**Question 3.** *It's just an implementation issue.* In effect that's right: All we have ever done is describe what a process memory can do, we've never built one! Situated cognition reformulates the knowledge level versus symbol level distinction. The knowledge level is an observer's specification of an agent-in-an-environment (cf. [5, 16, 17, 46]). Symbols aren't stored in memory and manipulated as structures; they are generated and interpreted anew with each perception. How memory is implemented (as opposed to described) is the essential problem we face as engineers. Memory-as-structure-storage models human intelligence, but can't replicate its flexibility.

**Question 4.** "*Why is it, although everybody now admits the force of the criticism of associationism, the associationist principles still hold their ground and are constantly employed?*" [4, p. 307]

> First, it is because the force of the rejection of associationism depends mainly upon the adoption of a functional point of view; but the attitude of analytic description is just as important within its own sphere. . . .
>
> Secondly, it is demonstrable that every situation, in perceiving, in imaging, in remembering, and in all constructive effort, possesses outstanding detail, and that in many cases of association the outstanding detail of one situation is taken directly out of that, and organised together with the outstanding detail of a different situation. . . .
>
> Thirdly, we have seen how to some extent images, and to a great extent words, both of them expressions often of associative tendencies, slip readily into habit series and conventional formations. They do this mainly in the interest of intercommunication within the social group, and in doing it they inevitably take upon themselves common characteristics which render them amenable to the general descriptive phrases of the traditional doctrines of association.
>
> In various senses, therefore, associationism is likely to remain, though its outlook is foreign to the demands of modern psychological science. It tells us something about the characteristics of associated details, when they are associated, but it explains nothing whatever of the activity of the conditions by which they are brought together. [4, p. 308].

**Question 5.** *How do you model medical diagnosis strategies now?* Just as we always have. Classifications and production rules are fine for stating behavioral patterns (what Bartlett calls "analytic description" of "habit series and conventional formations"). It remains to explain how they *develop* ("the conditions by which they are brought together"). Most learning programs grammatically describe how representations accumulate within a fixed language. They don't explain how representations are created, or more generally, the evolution of new routines not described by the given grammar.

**Question 6.** *Could there be a universal grammar?* Yes, in the sense that simple operations such as sequence, composition, and subsumption might constitute the "grammar" out of which all categorizations and behaviors are constructed. But these are processes, not descriptions or templates.

**Question 7.** *How would you test these theories psychologically?* In part, it's been done. Situated cognition research claims that empirical studies contradict cognitive science models. For example, Bartlett's [4] data argues against "fixed trace", semantic network memories. Jenkins [31] and Bransford et al. [9] demonstrate contextual effects that contradict models of memory based on search and matching. Other facts are obvious: people don't speak like grammatical automatons; emotions organize behavior; knowledge engineers are creating models—new representations—not extracting precoded networks from expert brains; AI programs aren't capable of writing reviews like this. The challenge goes the other way: What does AI have to say about Freud's analyses? About dreams? About musical ability? To pick a more mundane example, how could I write a page-long biography of a person, yet not remember his name? Schema-based storage models offer no explanation: How could I have access to a dozen feature-slots, but have no handle on the most obvious label for indexing the frame itself? Concerning psychological testing, note that situated cognition research rejects the validity of laboratory experiments that subtract out the complex, interactive context of everyday cognition [33]. Predictive social-psychological experiments are more like manipulating the weather than running rats through a maze.

**Question 8.** *Who is Rosenfield?* The book's jacket states,

> Israel Rosenfield received his M.D. from New York University and his Ph.D from Princeton. He teaches at the City University of New York and is the author of *Freud: Character and Consciousness* and co-author (with Edward Ziff) of *DNA for Beginners.*

For those swayed by authority, Oliver Sacks states in a recent article, ". . . I have been assisted by discussions with Pietro Corsi, Otto Creutzfeld, Gerald Edelman, Ralph Siegel - and, most especially, Israel Rosenfield" [50]. Rosenfield himself thanks researchers from the MIT AI Laboratory (regarding Marr's work), James McClelland (regarding PDP models), and others whose work he analyzes (Edelman, Marshall, Warrington).

**Question 9.** *It's all mystical.* Every science must exclude phenomena that are viewed as too complicated for a given paradigm. In cognitive science, problem-solving research has generally excluded emotion, psychiatric disorders, religious behavior, and other matters relating to personal identity and motivation. But it is sloppy thinking to contrast scientific methods (e.g., using logic and representational models) with ill-understood or apparently non-productive behaviors, thus circularly defining what needs to be explained by what can be modeled. We have done this in AI to the extent that logic is contrasted with emotion and deliberate reasoning with intuition. The very notion of "judgment" has become something expressible as a rule, when our every day experience is that judgments arise from non-represented thought. The currently popular view that common sense knowledge can be reduced to representations similarly distorts what needs to be explained.

The result is a science that resembles Oliver Sacks' patient, Dr. P., "the man who mistook his wife for a hat". Dr. P. apprehends objects by matching features of categories. For example, he describes a glove as "a continuous surface. . . infolded on itself. . . . [with] five outpouchings. . ." [49, p. 114]. Sacks comments,

> Our cognitive sciences are themselves suffering from an agnosia essentially similar to Dr. P.'s. Dr. P. may therefore serve as a warning and parable—of what happens to a science which eschews the judgmental, the particular, the personal, and becomes entirely abstract and computational. [49, p. 20]

## 11. Conclusions

What is memory? What is retained from experience? Memory is a capability to recompose sequences of behaviors, to coordinate past maps between perceptions and movements within a constructed context of ongoing perceptions and behaviors. In short, memory is indistinguishable from our capability to make sense, to learn a new skill, to compose something new. It is not a place where descriptions of what we have done or said before are stored. In more detail, memory-based performances involve an intricate combination of reconstructed "feelings" and "attitudes" that orient composition of new sequences, and specific reconstructed images, sounds, and other sensations that constrain behavior from "below". This is essentially Bartlett's model of constructive memory [4]; Rosenfield might have given him more credit for integrating emotion and sensation in this way.[16]

[16] The citation of Bartlett's work on p. 193 appears to be an afterthought. George Mead published a book two years after Bartlett's *Remembering*, with a strikingly similar emphasis on social aspects of cognition. For Bartlett and Mead, social interactions structure perception and meaning attribution. In this respect, we can criticize Rosenfield for not breaking severely enough with the egocentric view he means to attack. A more recent statement in this tradition is Lave's [33] eloquent discussion of how cultural knowledge cannot be reduced to representations that describe it.

Rosenfield's analysis provides a point of view for critiquing both cognitive science of the 1970s and 1980s and connectionism, while laying down requirements for a different model of perception and memory. To put it more strongly, as Rosenfield clearly intends, this book provides one starting point for overthrowing the views of memory that dominate the cognitive science and neural net community. Rosenfield sees both communities as ignoring basic properties of human memory that could be exploited in designing an intelligent machine: The prevalent cognitive science view is that memories are stored structures; the prevalent neural net view is that the world is received by the organism as an array of meaningful inputs (e.g., words).

It is no coincidence that Rosenfield cites case studies of patients who cannot read words or multinumeral digits. The question is not how we recognize or generate *single forms* or movements. But rather, how do we *coordinate a sequence* of perceptions or movements? This emphasis on coordination, correlation, and composition into coherent ongoing processes is the essential capability that connectionist research by and large has not yet addressed (but see, for example, [40]). A wonderful example of what situated cognition could bring to robotics is Ghengis' ability to develop the coordinated, tripod gait after a few minutes [36].

Situated cognition calls our attention to the environment and importance of on-going context. Our study of Rosenfield suggests a significant reformulation of the nature of context:

- the social-physical environment is important because behavior is perceptually organized by an interaction of environmental and internal processes;
- representations are created perceptually and interpreted in cycles of perceptual categorizing; in this they become part of the context that organizes behavior; to speak is to perceive is to represent is to learn;
- context is not given (i.e., as objective data); the context that causally directs perception is an *internal* construction of processes (neural sensory-effector maps and maps of maps), an ongoing composition;
- neural and perceptual theories are essential; the social sciences alone cannot explain why habits develop, account for the effects of practice, or most basically, explain how perceptual categories are biased by experience.

Recent interest in relating cognitive science to neurobiology, in what is called *cognitive neuroscience*, requires serious consideration of Rosenfield's attack. Looking for knowledge and stored plans in brain tissue is almost certainly fundamentally confused. Instead, we should take cognitive science theories as an observer's descriptions of what the brain, in interaction with an environment, accomplishes. We should look for a mechanism, along the lines of Edelman's neuronal group selection, which could account for how strategies and plans are manifested as compositions of sensory-effector maps, and how these orient the selection of neural groups, are modified by them, and endure.

There is reason to be optimistic that we are about to make some breakthroughs in our understanding of the brain. But ironically, it will come at the expense of overturning nearly everything AI has assumed about the *physical mechanisms* of perception, learning, and memory. Indeed, we must stop and rethink the rubric of "information-processing" that still unites most parts of AI and cognitive science. Perhaps our field betrays too much its origins in the computer industry, with data supplied on cards and each job completed as a neat processing from input to output piles. More generally, what is at stake is our ideas about how models relate to mechanisms and what mechanisms can be built. The most important clues remind us of powerful ideas from physics and biology: frames of reference, dynamics, development, and emergent structure. Having invented a kind of memory that ignores these aspects of life, we must now try to invent another.

## Acknowledgement

If memory serves me, extensive revisions of these representations were socially structured by Susan Chipman, Dan Russell, Steve Smoliar, Mark Stefik, and Kurt VanLehn. This work was supported in part by gifts from the Digital Equipment and Xerox Corporations.

## References

[1] P.E. Agre, Book Review of *Plans and Situated Actions: The Problem of Human-Machine Communication* (L.A. Suchman), *Artif. Intell.* **43** (1990) 369–384.

[2] J.R. Anderson and G.H. Bower, *Human Associative Memory* (Winston, Washington, DC, 1973).

[3] J. Bamberger and D.A. Schön, Learning as reflective conversation with materials: notes from work in progress, *Art Educ.* (March 1983).

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

[5] G. Bateson, *Steps to an Ecology of Mind* (Ballentine Books, New York, 1972).

[6] G. Bateson, *Mind and Nature: A Necessary Unity* (Bantam, New York, 1988).

[7] M.H. Bickhard and D.M. Richie, *On the Nature of Representation: A Case-Study of James Gibson's Theory of Perception* (Praeger, New York, 1983).

[8] P. Bordieu, *Outline of a Theory of Practice* (Cambridge University Press, Cambridge, 1932).

[9] J.D. Bransford, N.S. McCarrell, J.J. Franks and K.E. Nitsch, Toward unexplaining memory, in: R.E. Shaw and J.D. Bransford, eds., *Perceiving, Acting, and Knowing: Toward an Ecological Psychology* (Erlbaum, Hillsdale, NJ, 1977) 431–466.

[10] J. Bresnan and R.M. Kaplan, Grammars as mental representations of language, in: W. Kintsch, J.R. Miller and P. Polson, eds., *Method and Tactics in Cognitive Science* (Erlbaum, Hillsdale, NJ, 1984).

[11] R.A. Brooks, How to build complete creatures rather than isolated cognitive simulators, in: K. VanLehn, ed., *Architectures for Intelligence: The Twenty-Second Carnegie Symposium on Cognition* (Erlbaum, Hillsdale, NJ, 1991).

[12] J.S. Brown and K. VanLehn, Repair theory: a generative theory of bugs in procedural skills, *Cogn. Sci.* **4** (1980) 379–426.

[13] W.H. Calvin, *The River That Flows Uphill: A Journey from the Big Bang to the Big Brain* (Macmillan, New York, 1986).

[14] D. Chapman, Vision, instruction, and action, Ph.D. Thesis, AI Laboratory Tech. Report #1204, MIT, Cambridge, MA (1990).

[15] W.J. Clancey, Acquiring, representing, and evaluating a competence model of diagnosis, in: M.T.H. Chi, R. Claser and M.J. Farr, eds., *The Nature of Expertise* (Erlbaum, Hillsdale, NJ, 1988) 343–418.

[16] W.J. Clancey, The knowledge level reinterpreted: modeling how systems interact, *Mach. Learn.* **4** (1989) 287–293.

[17] W.J. Clancey, The frame of reference problem in the design of intelligent machines, in: K. VanLehn, ed., *Architectures for Intelligence* (Erlbaum, Hillsdale, 1991).

[18] W.J. Clancey, Why today's computers don't learn the way people do, in: P. Flasch and R. Meersman, eds., *Future Directions in Artificial Intelligence* (Elsevier, Amsterdam, 1991).

[19] W.J. Clancey, Model construction operators, *Artif. Intell.* (to appear).

[20] H. Cohen, How to draw three people in a botanical garden, in: *Proceedings AAAI-88*, St. Paul, MN (1988) 846–855.

[21] R.G. Collingwood, *The Principles of Art* (Oxford University Press, London, 1938).

[22] M. Cohtheart, Cognition and its disorders, review of *From Neuropsychology to Mental Structure*, *Science* **246** (1990) 827–828.

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

[24] S. Freud, *On Aphasia* (International Universities Press, New York, 1953).

[25] J.J. Gibson, *The Senses Considered as Perceptual Systems* (Houghton Mifflin, Boston, MA, 1966).

[26] R.F. Hadley, Connectionism, rule following, and symbol manipulation, in: *Proceedings AAAI-90*, Boston, MA (1990) 579–586.

[27] E.H. Hovy, Pragmatics and natural language generation, *Artif. Intell.* **43** (1990) 153–197.

[28] A. Iran-Nejad, Affect: a functional perspective, *Mind Behav.* **5** (1984) 279–310.

[29] A. Iran-Nejad, The schema: a long-term memory structure or a transient functional pattern, in: R.J. Tierney, P.L. Anders and J.N. Mitchell, eds., *Understanding Readers' Understanding: Theory and Practice* (Erlbaum, Hillsdale, 1987).

[30] W. James, *Psychology: Briefer Course* (Harvard University Press, Cambridge, MA, 1892); reprinted, with annotations (1984).

[31] J.J. Jenkins, Remember that old theory of memory? *Well, forget it! Am. Psychol.* (November, 1974) 785–795.

[32] B. Latour and S. Woolgar, *Laboratory Life: The Social Construction of Scientific Facts* (Sage, London, 1979).

[33] J. Lave, *Cognition in Practice* (Cambridge University Press, Cambridge, 1988).

[34] D. Lenat and R. Guha, *Building Large Knowledge Bases* (Addison-Wesley, Reading MA, 1990).

[35] E.F. Loftus, Leading questions and the eyewitness report, *Cogn. Psychol.* **7** (1975) 560–572.

[36] P. Maes, ed., Designing autonomous agents, *Rob. Autonomous Syst.* **6** (1, 2) (1990) Special Issue.

[37] G.H. Mead, *On Social Psychology* (University of Chicago Press, Chicago, IL, 1964); first published (1934).

[38] G.A. Miller and P.N. Johnson-Laird, *Language and Perception* (Harvard University Press, Cambridge, MA, 1976).

[39] M. Minsky, *The Society of Mind* (Simon and Schuster, New York, 1986).

[40] Y. Miyata, Organization of action sequences in motor learning: a connectionist approach, in: *Proceedings Ninth Annual Conference of the Cognitive Science Society*, Seattle, WA (1987) 496–507.

[41] U. Neisser, *Cognition and Reality: Principles and Implications of Cognitive Psychology* (Freeman, New York, 1976).

[42] D.A. Norman, *Learning and Memory* (Freeman, New York, 1982).

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

[44] J. Piaget, *Genetic Epistemology* (Norton and Company, New York, 1970).

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

[46] S.J. Rosenschein, Formal theories of knowledge in AI and robotics, Tech. Note 362, SRI, Menlo Park, CA (1985).

[47] S.J. Rosenschein, The logicist conception of knowledge is too narrow—but so is McDermott's, Tech. Note, SRI, Menlo Park, CA (1987).

[48] D.E. Rumelhart, J.L. McClelland and the PDP Research Group, eds., *Parallel Distributed Processing* (MIT Press, Cambridge, MA, 1986).

[49] O. Sacks, *The Man Who Mistook His Wife for a Hat* (Harper & Row, New York, 1987).

[50] O. Sacks, Neurology and the Soul, *The New York Review of Books* (November 22, 1990) 44–50.

[51] R.C. Schank, Failure-driven memory, *Cogn. Brain Theory* **4**(1) (1981) 41–60.

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

[53] D.A. Schön, *Educating the Reflective Practitioner* (Jossey-Bass, San Francisco, CA, 1987).

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

[55] L. Steels, Cooperation through self-organisation, in: Y. Demazeau and J.-P. Müller, eds., *Multi-agent Systems* (North-Holland, Amsterdam, 1989).

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

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

[58] L. Vygotsky, *Thought and Language* (A. Kozulin, ed.) (MIT Press, Cambridge, MA, 1986); original published (1934).

[59] M.M. Waldrop, Fast, cheap, and out of control, Research News, *Science* **248** (1990) 959–961.

[60] D.C. Wilkins, W.J. Clancey and B.G. Buchanan, On using and evaluating differential modeling in intelligent tutoring and apprentice learning systems, in: J. Psotka, D. Massey and S. Mutter, eds., *Intelligent Tutoring Systems: Lessons Learned* (Erlbaum, Hillsdale, NJ, 1988).

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

[62] D.A. Schön, The theory of inquiry: Dewey's legacy to education, Presented at Annual Meeting of American Educational Research Association, San Francisco, CA (1990).