ELSEVIER

Artificial Intelligence 101 (1998) 301–309

Artificial Intelligence

Book Review

Michael A. Arbib, *The Metaphorical Brain 2: Neural Networks and Beyond* ☆

John A. Barnden [1]

School of Computer Science, The University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom

Received 15 December 1995

1. Overview

This book presents a view both of computation in the brain and of how computation should be organized in future computer systems. As the book is written by someone who is well versed in several fields, including symbolic AI, artificial neural networks, neuroscience and control theory, it is well worth paying attention to. Much of the book is about matters such as vision, in both its low level and high level aspects, and movement control, but it also addresses natural language and, briefly, consciousness.

Apart from the particular theoretical view it espouses, the book is useful in its detailed summaries of some aspects of the mammalian brain, neural systems of lowly animals (notably frog, toad and marine slug), control theory, dynamical systems theory, artificial neural networks and AI. It is meant to be accessible overall to, say, a typical Scientific American reader, but is also intended to be useful for experts in fields such as AI, robotics, cognitive science, neural networks and neuroscience. The book is not, and could not be, anywhere near encyclopedic on such fields; the emphasis is rather on aspects of those fields that serve in some way to illustrate, support or constrain the central theoretical ideas of the volume. But one could repudiate the whole of Arbib's theory and still glean much useful information from the book, and indeed much of the information given seems rather independent of his theory. The book does skip about a bit from topic to topic, but as it is fat and dense that may be a good thing.

The central ideas of the volume are *cooperative computation, coordinated control programs, schemas, action-oriented perception, perception-oriented action*, and, stepping down to the neural level, *topological mappings* (in the sense that, say, a retinotopic representation in brain cortex is topological). In all this, schemas are the primary notion.

☆ Wiley-Interscience, New York, 1989.

[1] Email: J.A.Barnden@cs.bham.ac.uk; http://www.cs.bham.ac.uk/~jab.

PII: S0004-3702(98)00025-3

A schema is somewhat like a process description. Arbib views brain activity as a matter of many schema instances operating in parallel and communicating with each other. A schema instance is an active instantiation of a schema, and the population of instances can change rapidly. Two salient types of schema are perceptual schemas and motor schemas. Perceptual schemas are activated by, or look for, particular things in the perceived environment. These things might be relatively low level features or they might be members of some high-level conceptual category. A perceptual schema instance is not only a process but also holds information about the entity it has perceived (so to speak). Motor schemas are prescriptions for motor activity, and are a special case of schemas for action.

Arbib typically uses the term "cooperative computation" to mean the interaction of schema instances. But he also uses it (at a coarse grain) to describe the interaction of brain modules that he does not explicitly classify as schema instances, and (at a fine grain) to describe interactions between units in a neural network, where again the units are not cast as schema instances.

In Arbib's view, action can be perception-oriented in that an action can be aimed at facilitating further perception. Moving the eyes or head is a basic example of this. Conversely, perception is action-oriented in that perceptual schemas are aimed at facilitating specific actions the organism may take. Importantly, however, those actions can be internal ones that are aimed at updating or improving the organism's model of its world, rather than being external actions on the environment. The internal model has a short-term aspect consisting of schema instances, and a long-term aspect consisting of schemas.

One way in which the action-oriented quality of perception is important in Arbib's view is that low-level perceptual computation might be enough for various types of action. For instance, in order to navigate through some objects it can be enough to know where the objects are, not what they are. Relatedly, Arbib presents evidence that the brain has various different vision systems, with object locating and object recognizing being done by different systems.

Schema instances can be grouped into schema assemblages, which can themselves be instances of higher-level schemas. (One minor problem I had with reading the book was the use of the term "schema assemblage" when "schema instance assemblage" would have been clearer.) The dynamic formation of such higher-level schemas is one type of learning Arbib hints at, although he does not provide a detailed account of how it might work. A coordinated control program is a schema whose instances are schema(-instance) assemblages, and which is partly like a flowchart and partly like a system chart of the type used in control theory. The nodes are schemas that can communicate both by activating each other and by continuously passing signals to each other. The amalgamation of control theory with the types of interaction covered in the parallel computing sector of computer science is one of the more distinctive features of Arbib's account.

Arbib occasionally appeals to the Piagetian notions of assimilation and accommodation (to and of schemas, respectively). However, it was not clear to me how his schemas resemble or differ from the schemas of Piaget, or the similar constructs of other authors whose work he reviews. (He does point out that his work differs from that of Neisser and Gibson in being more concerned with the detailed *mechanisms* underlying cognition—p. 47.) In particular, although Arbib mentions AI frame-based and script-based research as

being part of the same enterprise as his schema theory (p. 47), and briefly points out some differences, I would have appreciated a more detailed comparison. Arbib mentions that in this book he is not concerned with the "social schemas" he has discussed elsewhere. These are patterns of behavior involving several organisms, and are held by the society en masse (p. 45). The book confines itself to schemas inside individual organisms.

A strong methodological theme in the book is that it is important to have tools for describing computation at levels higher than that of neural networks, and of course he proposes schema theory as an important item in the toolbox. At the same time he is careful not to make the simplistic claim that a schema, schema instance, or a component of a schema or instance is implemented in the brain as a dedicated neural network. He points out that a single region of the brain might subserve many schemas, and a schema (or instance) might involve several regions. (I would have preferred less talk of regions as opposed to just subsets of the total set of neurons.) In particular, he guards against assuming that, just because a neural network exhibits two different behaviors under different conditions, it must have a subnetwork set aside for each behavior. He gives a simple example of this (p. 11): a small network that can either act as an AND gate or as an OR gate depending on the value on one input line, but where there is no division into a module for AND and one for OR.

In this way Arbib is sympathetic to high-level units of behavior emerging in some subtle, distributed way from neural networks. This resonates with standard connectionist claims that such things as rules can be merely emergent from subsymbolic computation. On the other hand, Arbib also says that high-level units such as rules can enable the system to avoid a great deal of expensive, highly parallel, low-level processing (p. 247). I found this balanced view of Arbib's appealing, but the claim about high-level rules needs further justification, as one might have conjectured that highly parallel, low-level processing is not a drain on the overall system in the first place.

Another interesting claim of Arbib's on the subject of levels is that the brain develops high-level motor schemas and also *interpreters* for these schemas, one such for many different body systems. Various different body systems may therefore be capable of interpreting a schema (with a possible degradation in performance) even though they have never seen that particular schema before, (p. 251). This idea is useful in accounting, for instance, for the fact that you can use a paintbrush tied to a long pole to write your name on a wall, in a somewhat individual way. The schema interpretation notion is also used in a theory of hand movements couched in terms of "virtual fingers." A virtual finger can be mapped to a subset of the fingers (including the thumb), or even to the palm of the hand. High-level schemas are expressed in terms of virtual fingers, leaving context-sensitive interpreters to map virtual fingers to particular sets of hand parts. The primary example here is the act of picking up a mug, where the particular set of fingers put inside the handle depends on various features of the mug.

Despite Arbib's emphasis on distributed computation in general and certain types of distributed effect in neural networks, some readers may feel that his approach leads to too discrete a division of behavior into schemas. For instance, in discussing prey acquisition by toads, Arbib proposes (pp. 219–222) one perceptual schema for detecting a barrier, one for detecting a chasm, and one for detecting the free prey condition (no obstacle, i.e., barrier or chasm). But there is no argument that it is indeed appropriate to postulate separate

schemas here. Perhaps there is one schema from which different behaviors can emerge (cf. the AND/OR example above). Arbib may fact be right in his analysis of prey acquisition by toads, but it is not clear that the sort of division he proposes here would hold water for other activities of other organisms.

As Arbib himself points out, he is not presenting a particular formalism for schemas. One is left to induce what they are like from the examples he gives—and he does indeed give some detailed examples. Arbib rightly points out that in computer science the general notion of a program (as opposed to a program in a specific language or theoretical framework) is, equally, only vaguely specified; but it is still a useful notion.

Arbib also freely admits that he does not have an account of how multiple, dynamically arising instances of a given schema can be handled in neural circuitry, whether artificial or biological (p. 246; see also p. 180). I concur with him that this is a fundamental and crucial issue, and I have made brief comments on it elsewhere (Barnden [4]). As Arbib says, one central technical issue is the problem of how inferential and other mechanisms respond to recruited, and therefore in someone sense unpredictable, assemblies of neurons. (This is on the presumption that a schema instance is realized as a recruited assembly, though that is not the only possibility). Actually, the more fully worked out applications of schema theory do not involve multiple instantiation—for example, the perceptual schema for barriers in toads is its own unique and permanently existing instance. But multiple instantiation is multiply appealed to elsewhere in the book.

Schema theory is, overall, rather speculative, though in places backed up by experimental data or by plausible connections with neurophysiology. Much of the book does not provide (and does not appear to be aimed at providing) evidence that his theory is correct, and instead supplies interesting constraints or boundary conditions on what a correct, fully developed schema theory would be like. For instance, the detailed discussion of the dynamics and feedback control of muscles makes little direct contact with schema theory (though clarifying what motor schemas need to do in order to control muscles), and the same is true of the overviews of artificial neural net frameworks. The detailed accounts of some aspects of human brain circuitry are at best suggestive of how schemas might be neurally realized. But all this is not surprising or reprehensible, and as Arbib points out, speculations about how schemas could be realized in neural networks, or about what brain networks they might involve, can prompt useful experiment and can lead to new developments in neural net research (p. 207).

The book's title alludes to metaphor. This is for two reasons. First, Arbib views any scientific theory as a metaphor for its target, and views his schema theory in particular as a metaphor for the brain. As part of this, he alludes to the metaphor of the brain as computer (but not as a traditional serial one). In a sense Arbib inverts the brain-as-computer metaphor in suggesting in the final chapter that future computers will compute in the style of the brain as construed through schema theory. Some of his specific suggestions on this topic are summarized below. The second reason for the title is that he casts the brain's schemas as metaphorical constructions of chunks of reality. In addition, Arbib mentions his view that language is inherently metaphorical (endnote 7 in Chapter 5).

Finally, the book is a radical revision of his earlier book with a similar title (Arbib [1]). Some technical concepts have been renamed and elaborated—in particular, the so-called slide-box metaphor of the 1972 book is superseded while still having some usefulness

(pp. 32–34). Also, much extra material has been added. The present book does not rely on familiarity with the earlier one.

The following section discusses Arbib's schemas further. The section after that briefly outlines what Arbib covers in various areas: AI, sixth-generation computing systems, high-level cognition (especially language), consciousness, and neural networks. (The extensive material on the structure and wiring of the brain and related organs is not systematically outlined here.) The final section is a summary of the review.

2. Some issues concerning schemas

Arbib's schema theory would, I imagine, be appealing to many people in AI, particularly as Arbib gives a detailed account of how the knowledge-based VISIONS system for computer vision illustrates the theory, and also discusses in more general terms how blackboard systems such as HEARSAY II relate to the theory. Also, Arbib presents some schema-theoretic work on manipulator control in robotics, and summarizes schema-theoretic accounts of various aspects of natural language (see below). And, schema theory is in tune with the current emphasis on parallel and distributed computing.

However, some aspects of Arbib's account are not entirely convincing. Arbib suggests that in applying schema theory to understanding images of complex scenes, different schema instances would be brought into being by a given portion of the image, and they would compete with each other to provide an explanation of that portion. For example, one part of the image might stimulate the creation of both an instance of a tree-foliage schema and an instance of a blob-of-ice-cream schema; similarly, an adjacent portion might stimulate instances of a tree-trunk schema and an ice-cream-cone schema. (See pp. 37–39 and the discussion of the VISIONS system.) The ice-cream and foliage instances would compete with each other, while the foliage instance would cooperate with the trunk instance, and so on. This type of suggestion is natural and quite common, especially in connectionist circles. However, one needs to get to grips with what would happen to the proposal if it were scaled up to reality. The sort of image region that could suggest either ice-cream or foliage would, in a full cognitive system, equally suggest a myriad other categories (a fluff of cotton, a cloud in the sky, an cloud of car exhaust, a blob of mashed potato, a hairdo, a smudge on a window-pane, etc. etc.) One would therefore have to have a competition/cooperation scheme that would be able *rapidly* to deal with very large numbers of schema instances. While it is conceivable that existing computational schemes for competition/cooperation (e.g., in neural networks) could handle cases of realistic size, this needs to be demonstrated. To be fair, Arbib does essentially point this problem out (p. 264) in discussing the VISIONS system.

I felt there should have been some consideration of the possibility that overall collections of low level image features somehow stimulate schema instances characterizing the whole image (or large parts of it), without first stimulating schema instances for smaller parts of it; and those large-scale schema instances then suggest schemas for smaller parts. I am not claiming that this proposal has a better future than Arbib's own, but only that a wider range of possibilities needs to be considered.

The book consistently views multiple instances of a given schema as individually active entities—simultaneously and independently running "invocations" [my term] of the schema, which is merely their "program." However, readers should be aware that at a recent workshop on schema theory (Arbib et al. [2]), Arbib voiced the possibility that in many case it might be better to propose that instances are merely repositories of particular parameter settings for the schema, and not simultaneously active; in this view, the schema would be the only thing that would be active, and at any time would use the parameter settings provided by some single instance. (See footnote 3 in Arbib's Introduction in Arbib et al. [2], and also another paper by Arbib in the same proceedings.) This would seem to be a major shift of viewpoint, but the general consequences for schema theory remain to be spelled out.

At one point Arbib is not entirely true to his own precepts on action-oriented perception. In describing the VISIONS system as one that conforms to his schema theory, Arbib discusses the fact that the system failed to see a mailbox in the input scene, and also failed to identify a white patch behind a tree as part of the wall of a house it had identified (p. 244). Arbib discusses modifications to the system that would allow it to spontaneously notice these things. However, I had not in fact noticed the mailbox in the picture, or realized that the white patch was part of the house! This was so even despite my close examination of the picture while reading the previous text. I only perceived the mailbox and patch for what they were in executing the action of reading Arbib's discussion of them.

Finally, one particular expansion of Arbib's view of schema-instance interaction may be appropriate. Arbib gives the impression much of the time that instances interact with each other by rather simple signals or messages. As a major exception, in the VISIONS system the instances interact by setting up goals and hypotheses visible to the system at large. Clearly, for high-level cognitive in general it is appropriate to imagine schemas communicating by means of rather complex messages. It would seem to me fruitful to cast complex messages themselves as schema instances. For example, the just-mentioned goals and hypotheses could be schema instances.

3. Outlines of area coverage

On *artificial intelligence*, Arbib spends most of his time discussing vision systems, particularly the knowledge-based VISIONS system. There is some detail on methods for determining the depth of objects away from a vision system, including but not limited to stereopsis, and of the determination of optic flow. Arbib casts doubt on the need for, or realistic feasibility of, the $2\frac{1}{2}$-D sketches of Marr (p. 224f, 363). Arbib intends schema theory in general, and particular tools for hand control such as virtual fingers and opposition spaces, to provide a framework for robotics. However, there is not much discussion of robotics as such, although he does briefly describe a schema-based path-planning system, AURA, and a neural network landmark-learning system (see below). (For more work on application of schema control to robotics, see the above-referenced workshop proceedings.) Planning of actions is often mentioned in the book, though no detailed planning scheme is set out. The virtual-finger mechanism amounts to a specialized type of planning, and is described in a relatively detailed way. The HEARSAY II

speech-understanding system is overviewed, as it has some kinship with schema-theory (particularly because VISIONS is reminiscent of HEARSAY II).

Late in the book (p. 392) Arbib throws out a large claim merely in passing. He says that "... schemas ... determine a course of action by a process of analogy formation, planning, and schema interaction, which need have little in common with formal deduction." (Readers should realize that it is schema instances that are really meant here, not schemas as such.) My focus here is on the mention of *analogy*. No other mention of it occurs in the book, except implicitly through the rather general comments on metaphor (as summarized above) and Piagetian assimilation. I believe that analogical processing as a replacement for more rigid deductive reasoning is an important and fruitful idea, one that could usefully have received extensive and detailed discussion in the book.

Arbib outlines three schema-theoretic models in the realm of *natural language*: a lesionable cooperative computation model of sentence comprehension (developed by H.M. Gigley); a model of language acquisition in the two-year-old (developed by J.C. Hill); and a model using salience in deep generation of natural language descriptions of visual scenes (developed by E.J. Conklin). These models are described in much greater detail in Arbib, Conklin & Hill [3]. (Only the first of the models is anywhere near a neural net level of description.) In discussing language acquisition, Arbib repudiates the notion of innate lexical classes or universal grammar. Instead, he seeks to account for acquisition partly by appealing to children's goal to communicate and their liking for repeating utterances back. Words that are used in similar ways come to be assigned to the same class, and the model assumes innate schemas for usage-based classification. There are also some other innate schemas. Arbib says that learning the grammar of a language is a very small part of learning the language—the child must also master the idiosyncratic morphophonology of the language, the subtle meanings and interrelations of a huge vocabulary, and a large stock of idioms, phrases and metaphors.

Arbib's general view of language is summed up in an endnote (p. 412): "...meaning is extracted from a virtually endless dynamic process. A sequence of words is always an impoverished representation of some schema assemblage. A sentence may have a literal meaning in terms of skimming off the most common set of associations from the related schemas, but there is no dividing line that sets off the literal from the metaphorical. In all cases, the schema or discourse provides an entry into a schema network." Also, Arbib sees language as being rooted in sensorimotor processes (p. 264), and claims (p. 181) that the overall similarities between sensorimotor and language computations are such that we should be able to find clues to language mechanisms by studying sensorimotor activity.

Arbib briefly describes a preliminary schema-based theory of *consciousness*, drawing in part on the evolutionary ideas of Hughlings Jackson. To communicate, an organism must be able to form a summary abstracted from internal neural complexities. Such summaries evolved to become a basis not only for external communication but also for internal planning and coordination. It is the activity of this co-evolved process that constitutes consciousness. Arbib also views communication as just a natural progression from body control: "there is a continuity from controlling one's own body, to using tools, to using another member of one's group to complete some action. As far as the brain is concerned, there is no self that stops at the end of the fingers." (p. 392.)

Arbib is sympathetic to the use of *imagery* as a tool in inference (pp. 226–229). While he devotes most of his discussion to the Shepard and Metzler experiments on imagined 3D object rotation and reflection, I was more interested in the suggestion that a system could answer the query "Is President Bush a man" by accessing an image of him and applying a pattern-recognition routine to it. (Arbib merely uses this as a simple, suggestive example of a general style of reasoning—he is not seriously suggesting that this particular query should be so answered.) Other researchers have made similar suggestions, both within AI and without, but it has yet to become a central proposal within AI.

At the end of the book, Arbib devotes some time to presenting his view of *sixth-generation computing systems*, whether for AI or not. He foresees systems that are richly embedded in their environments, blur the distinction between computers and robots, can communicate with users through graphical means, are capable of learning, may have neural networks as subcomponents, and are organized overall as a collection of cooperating agents. Much of this is, of course, in tune with current work on distributed computer systems, interfaces, and so forth. He sees programming as still being important, but not so much in the present-day sense as in the sense of defining the task specifications of small agents, which would then learn to do the tasks.

Arbib provides descriptions of various well-known types of *neural net*, notably: perceptrons; an associative network of Barto and Sutton for learning how to navigate with respect to some landmarks; Amari & Arbib winner-take-all networks; Hopfield nets; Boltzmann machines; and back-propagation nets. The inclusion of a detailed mathematical presentation of a winner-take-all network was welcome, because mechanisms for winner-take-all are given short shrift in some introductory texts on neural networks. On the other hand, I noted the lack of references to other methods for achieving winner-take-all behavior. I was also surprised that there was no description of Grossberg's ART systems (based on Adaptive Resonance Theory), as it seems relevant to a discussion of model matching (p. 375). And, despite Arbib's emphasis on topological maps in the brain, there was no description of systems for automatic learning of topological maps, but only a reference to the relevant works by Kohonen and Arbib.

Arbib is somewhat sceptical of the promise of existing neural network learning methods to provide an account of human learning as a whole (p. 373–375, 389, 391). He feels that declarative memory is not well served, and that domain-specific learning techniques will be needed.

The book provides a lot of interesting detail about neural networks in and overall organization of the *central nervous systems* of humans, other mammals, toads, frogs and marine slugs (Aplysia). There is useful discussion of oscillatory circuits, muscle-control circuitry, and gaze-control circuitry (including the role of retinotopic maps). There is a great deal of detail on the interconnectivity and possible roles of the many visual, motor and oculomotor regions of the mammalian brain. He speculates usefully on the brain regions that could substantiate his virtual-finger theory, although he does not suggest circuitry that could dynamically assign virtual-fingers to real hand-parts. He provides a useful caveat about the interpretation of Lashley's experiments on maze learning by rats, often quoted as showing a high degree of distributedness of function in the brain (cf. Lashley's laws of mass action and equipotentiality). Arbib mentions a repetition of Lashley's experiments, in which, instead of lesion-induced impairment of maze-learning

performance being measured by one overall parameter, impairment was differentiated into several types. For example, some lesioned rats had a tendency to turn left, others to still, others to be easily distractable. Thus, it seems that specific types of lesion cause specific types of impairment.

4. Summary

The book is thought-provoking and informative, wide in scope while also being technically detailed, and still relevant to modern AI even though it was published in 1989. This relevance lies mainly in the book's advocacy of distributed computation at multiple levels of description, its combining of neural networks and other techniques, its emphasis on the interplay between action and perception, and its particular approach to natural language processing. The criticisms I have occasionally made above are relatively minor, and one of the major shortcomings of schema theory at least at the time of the book's publication—namely the lack of a detailed, preferably neural, mechanism for schema instantiation—is one of which Arbib is well aware. The book is well produced, for the most part clearly written, and has useful summaries at the beginning of each chapter. Both general readers and AI researchers (mainly but not exclusively those working on robotics, vision and neural networks) could benefit from reading the book.

References

[1] M.A. Arbib, The Metaphorical Brain: An Introduction to Cybernetics as Artificial Intelligence and Brain Theory, Wiley-Interscience, New York, 1972.

[2] M.A. Arbib, G. Bekey, D. Lyons and R. Sun, Proceedings Workshop on Neural Architectures and Distributed AI: From Schema Assemblages to Neural Networks, Technical Report, Center for Neural Engineering, University of Southern California, Los Angeles, CA, 1993.

[3] M.A. Arbib, E.J. Conklin and J.C. Hill, From Schema Theory to Language, Oxford University Press, 1987.

[4] J.A. Barnden, The centrality of instantiations, Behavioral and Brain Sciences 10 (3) (1987) 437–438. (Commentary on a paper by Arbib in same issue.)