Getting the Ghost out of the Machine: A Review of Arnold Trehub's The Cognitive Brain

Luciano da Fontoura Costa
Cybernetic Vision Research Group
IFSC-USP, Caixa Postal 369
Sao Carlos, SP, 13560-970
BRAZIL

luciano@ifqsc.sc.usp.br

Received: August 19, 1994; Accepted: November 2, 1994

PSYCHE, 1(15), January 1995
http://psyche.cs.monash.edu.au/v1/psyche-1-15-dacosta.html

Keywords: cognition, consciousness, computational models of cognition, motivation, retinoids.

Review of: Arnold Trehub (1991) *The Cognitive Brain.* Cambridge, Mass.: MIT Press. 342 pp. Price: $27.50 pbk. ISBN 0-262-20085-6.

> "The next generation of computers will be so intelligent that we will be lucky if they keep us around the house as household pets." (attributed to M. Minsky in Searle (1989)).

1. Introduction

1.1 *The Cognitive Brain*, by Arnold Trehub, is an unusually comprehensive and enlightened attempt to develop a theoretical framework for understanding cognition. Based mainly upon two basic neural structures, the synaptic matrix and the retinoid, Trehub develops a series of models for many of the most important functions of the brain, including pattern recognition, selective attention, semantic processing, motivation and planning. The book also describes how these models have been successfully implemented in computers and how they are supported by neurophysiological, psychophysical and clinical evidence.

1.2 It comes as a surprise to find that a book dealing so comprehensively with cognition fails to address explicitly the mind-brain problem or the philosophical problem of

consciousness generally -- in fact, the term 'consciousness' does not even appear in the index, and the term 'mind' appears with but one reference. However, we can find a few passages treating such issues, such as one at the very end of the book: "It is the total specific content, the current physical state of specialized mechanisms in an individual brain shaped by encounters in a world both real and imagined, that constitutes the mind" (p. 305). Such a statement, together with the underlying philosophy adopted throughout the book, suggests that Trehub may be a constructive naturalist (see Flanagan, 1992), one who particularly believes that "we now have sufficient knowledge of the physiology of nerve cells and the structure of the brain to advance the theoretical formulation of putative brain mechanisms that can account for the basic competence of human cognition" (p. 2).

1.3 I begin by briefly looking at some of the principal developments in Trehub's book and follow by continuing from where Trehub leaves off -- i.e., by discussing the implications of Trehub's discussion for consciousness research.

2. A Brief Review

2.1 One of the main themes of Trehub's book is that the levels of explanation for cognitive phenomena -- put forward by David Marr (1982) as levels of computation, algorithm and physical implementation -- are *interdependent*, rather than independent. Thus, computational models and theories of the cognitive processes should be developed according to constraints imposed by the algorithmic level which, in turn, should be constrained by the implementation level. The obvious advantage of such an approach is to reduce the number of alternatives that can be considered for modelling cognition.

2.2 Accordingly, one of the first steps pursued in the book consists in specifying the assumed basic biophysical properties of neurons, especially the mechanisms through which long- and short-term learning and memory can be achieved, which in turn constrain the universe of possible cognitive models. The dynamics of the long-term mechanisms are assumed to be controlled by the density of the axon transfer factor (ATF) and dendrite transfer factor (DTF), which leads to the reinforcement of those synaptic contacts presenting a reasonable degree of coactivation between pre- and post-synaptic activity (a kind of Hebbian learning). Short-term storage is assumed to be implemented through autaptic cells, a special type of neuron that incorporates synaptical positive feedback through recurrent collaterals of its own axons.

2.3 Those are the basic elements which constitute Trehub's two principal types of neural structures: the synaptic matrix and the retinoid. The former (overall structure given in Figure 1) includes two sub-matrices, namely the imaging matrix and the detection matrix, which are interconnected. The adaptive synapses are capable of long-term learning provided by the dynamic interaction between ATF and DTF. Synaptic matrices receive input at the imaging matrix and produces output from the detection matrix through filter (f) and class (O) cells. Filter cells implement the integration of the input signals relayed by the imaging matrix via the mosaic cells; class cells receive inhibitory input feedback from their own outputs via the reset cell, which is needed in order to implement a kind of

`winner-takes-all' mechanism ensuring that only one class cell will activate for each specific input pattern, thus implying that specific stimuli be coded in terms of the relative spike frequency of the class cells. The output lines are feedback into the imaging matrix as a means of implementing associative recall. The synaptic matrix supports not only the ability to learn and recall static semantic representations (associative memory) but also to learn episodic representations, i.e. sequences of symbols temporally correlated (both in vision and in audition). More elaborate cognitive processing can be achieved by interconnecting synaptic matrices.

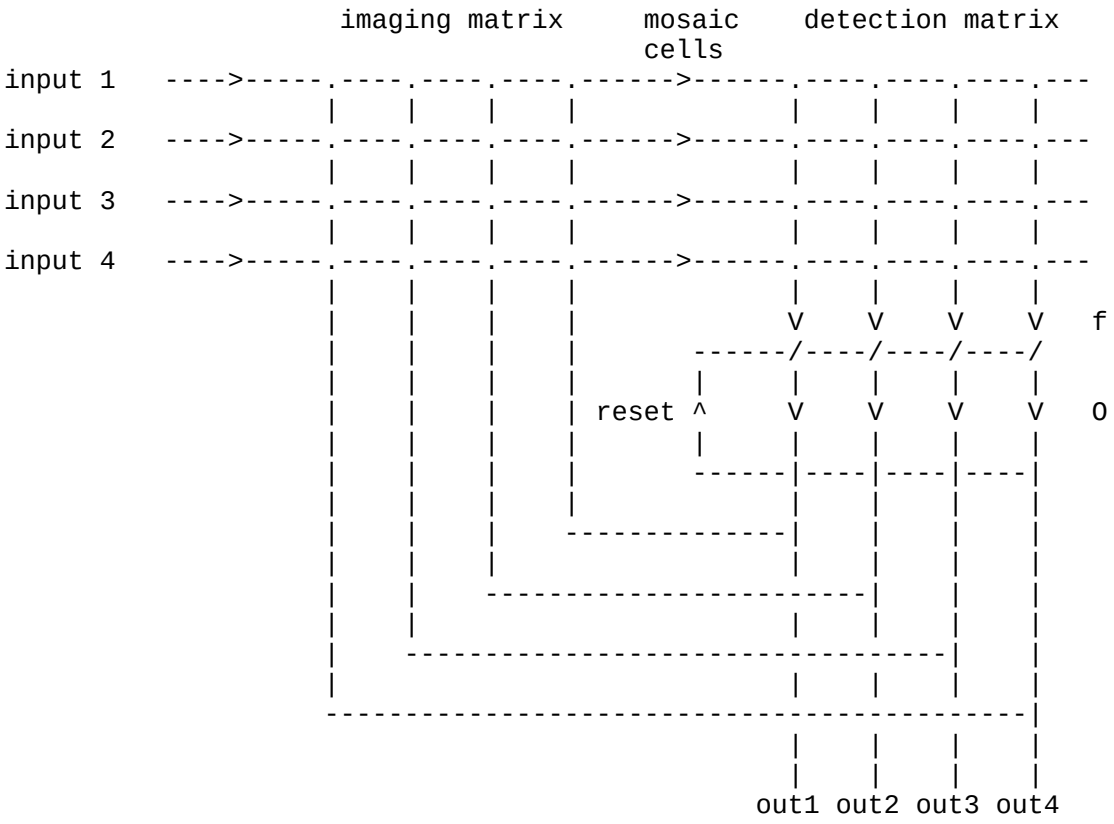

Figure 1: General structure of a synaptic matrix. "V, ^, >" = neurons; "." = adaptive synapse and "/" = inhibitory synapse.

2.4 The other basic neural structure in Trehub's approach is the retinoid, a dynamic post-retinal buffer built up of retinotopically organized autaptic neurons providing short-terms storage. Typically, retinoids receive topographical projections from other visual processing and input modules. Retinoids are combined in the retinoid system, which is capable of performing a series of cognitive tasks including: (i) parsing objects; (ii) constructing 3-D representations; (iii) locating and representing the self with respect to a specific environment; (iv) representing paths of movement and; (v) implementing selective attention. An important aspect of the retinoids is that they are organized around the normal foveal axis, which provides a central point for spatial reference. Selective attention, an important perceptual mechanism incorporated into the retinoid system, is achieved via a special reference retinoid, namely the self-locus retinoid, which may

"point" to different positions in visual space. Additional mechanisms are incorporated into the system in order to provide stereo visual processing and representations invariant across geometrical transformations -- these latter being invaluable for keeping the number of representations of recognizable elements manageable.

2.5 By using retinoids and synaptic matrices as well as elaborated versions of these structures, Trehub attacks a number of important cognitive processes, including planning, analysis of object relations, composing behaviour, motivation, character recognition, self-directed learning and narrative comprehension. Although a comprehensive discussion of these mechanisms is out of the scope of the present review, some additional remarks are due on specific consciousness-related issues. Identified as the "core function among the executive processes" (p. 295), the cognitive process of motivation is one such. Trehub's analysis of motivation is based upon the hedonic centre, which includes two homeostatic subsystems, designated HS-I (exclusively related to internal processes) and HS-II (related to external, i.e. worldly or "secular" processes). The dynamics of the homeostatic subsystems determine the "pleasure" or "displeasure" of the system; for instance, experimental findings seem to indicate that, upon deviation from the set point (equilibrium) of one of the homeostatic subsystems, suitable actions directed to the restoration of the equilibrium can induce pleasure. The hedonic centre is thus responsible for controlling the individual's actions aimed at the fulfilment of priority goals, which can be achieved through planning (performed by a synaptic matrix system) and action. Another important consciousness-related issue in Trehub's discussion concerns the I-tokens. These are special kinds of autaptic neurons which derive their input from the self-locus retinoid, yielding, when properly interconnected with other active cells, a state of "personal belief." The importance of such special neurons can be immediately appreciated from Trehub's own words: "I-activated predicate tokens, in turn, can evoke sensory images by their chain of backward links to the imaging matrix in a level-1 synaptic matrix. The extended set of such neuronal associations can be taken as the biological substrate for one's sense of self" (p. 302).

3. When Saying a Little is Saying a Lot

3.1 Perhaps the main contribution of Trehub's book to consciousness studies consists precisely in the fact that it has avoided such issues altogether. Indeed, by developing and implementing effective and biologically plausible computational models for many of the cognitive processes in the primate brain, *The Cognitive Brain* suggests that machines possessing abilities similar to humans *can* be implemented without worrying about any of the issues that makes consciousness research controversial (typically, feelings, experiences or qualia), a position that can be traced back to T. Huxley (Flanagan, 1992) and which has been recently revived (e.g. Korb, 1991, and Chalmers, in press). Although limited in scope, the computational models reported in Trehub's book nevertheless must support some optimism about the possibility of developing machines that could not easily be distinguished from humans. Now, whether such machines would then be acknowledged as conscious or not is a moot question, one that strictly depends upon the definition of 'consciousness' adopted. If we assume that being conscious is to know about oneself as well as one's interaction with the rest of the world, results like those presented

by Trehub substantiate the feasibility of conscious machines -- which appears very much in agreement with the positions expressed in (Korb, 1991) and (Chalmers, in press). In fact, the principal mechanisms for implementing such properties are outlined by Trehub and consist, roughly, in the interaction between the central hedonic centre (providing motivation for actions), the self-locus retinoid (providing information about one's spatial position), and stored knowledge about the individual, the rest of the world, and how these two can interact. It is interesting to observe that, if we concede that consciousness has a graded nature (as seems intuitively right) and if we adopt the definition of 'consciousness' suggested above, then it should be acknowledged that some currently available robots -- those that can accomplish tasks involving visuomotor integration -- already are conscious, to a limited degree.

3.2 As pointed out by Chalmers (in press), the really tricky issue usually related to consciousness is whether such machines would somehow experience sensations as humans do. Trehub has chosen to avoid such problems altogether, as is clear from his own words: "I must, therefore, acknowledge at the outset that the models I present do not aim to explain directly such ineffable matters as the felt qualities of a breathtaking sunset..." (p. 2). A wise decision, since our current knowledge seems to me too incomplete to support effective scientific work on the issue, although Chalmers (in press) makes a brave attempt in that direction. It should however be emphasized that this does not commit me to dualism or any other sort of mysterianism; my claim, evidently shared by Trehub, is merely that we are not yet in a position to mount an assault on this central philosophical problem of consciousness. No such pessimistic reflections should bother those pragmatic people who, for instance, are interested in understanding many of the neuroscientific *aspects* of consciousness or who are developing versatile and effective machines that can interact naturally with humans.

4. Concluding Remarks

4.1 Trehub's book consists in a comprehensive and enlightened approach to computational modelling of cognitive processes in the brain; and in this capacity it deserves greater attention than it has received thus far. Although *The Cognitive Brain* presents models for many consciousness-related issues, it does not explicitly address the mind-brain problem or the problem of qualia. The successes in cognitive modelling which Trehub recounts, however, strongly suggest that the pragmatic goal of building machines which can interact naturally and intelligently with humans may be achievable in the not too distant future. Indeed, if we define 'consciousness' so as to leave out qualia or feelings, while retaining what most neuroscientists and psychologists studying such phenomena have found interesting (such as attention, planning and a knowledge of self), we appear to have every reason to believe that consciousness research is ready to make substantial and direct progress. In other words, the disputes that form the core issues for the philosophical problem of consciousness may well be only peripheral issues for the scientific problem of consciousness. An examination of Trehub's book may be useful in assessing such a perspective.

Acknowledgements

The author would like to express his gratitude to Prof. C. Koch (Caltech), Dr D. Chalmers (Washington University) and Dr. K. Korb (Monash University) for their help in obtaining some of the references and to Prof. R. Koberle (IFSC-USP) for useful discussions.

References

Chalmers, D. J. (in print). Facing up to the problem of consciousness. In *Proceedings: Toward a scientific basis for consciousness* Cambridge, Mass.: MIT Press.

Flanagan, O. (1992). *Consciousness reconsidered*. Cambridge, MA: MIT Press. Book review available

Korb, K. (1991). Searle's AI program. *Journal of Experimental and Theoretical AI, 3*, 283-296.

Marr, D. (1982). *Vision*. San Francisco: W. H. Freeman.

Searle, J. (1989). Minds and brains without programs. In C. Blakemore & S. Greenfield (Eds.), *Mindwaves*. Oxford: Basil Blackwell.

Trehub, A. (1991). *The cognitive brain*. Cambridge, MA: MIT Press.