Do Seated Souls Experience Slumberous Sensations?

Review of *The Engine of Reason, The Seat of the Soul* by Paul Churchland

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

luciano@ifqsc.sc.usp.br

Received: November 21, 1995; Accepted: April 18, 1996; Published: June 18, 1996

PSYCHE, 2(29), June 1996
http://psyche.cs.monash.edu.au/v2/psyche-2-29-dacosta.html

KEYWORDS: neurocomputing, consciousness, brain research, recurrent networks, vector coding

REVIEW OF: Paul Churchland (1995) *The Engine of Reason, The Seat of the Soul.* A Bradford Book/The MIT Press. i-xii + 330 pp., 89 illus. Price: $US29.95 pbk. ISBN: 0-262-03224-4.

1. Introduction

> 'I feel, therefore I can appreciate ice cream.' (Folk Psychology)
>
> 'My personal computer has recurrent networks, therefore it can appreciate ice cream.' (New Folk Psychology)

The Engine of Reason, The Seat of the Soul (Churchland, 1995), EORSOS for short, is composed of two main parts. Whereas the first part presents a well motivated and accessible introduction to important issues in neurocomputing, the following part mainly dwells on controversial philosophical issues concerning consciousness, morality and society. The following sections present a critical overview of Paul Churchland's new book.

2. A Tasteful Introduction to Neurocomputing

One of the main purposes of the first part of EORSOS is to acquaint the general reader with the emerging area of neurocomputation, and especially with the concepts of vector coding, vector completion and recurrent neural networks, which play an important role throughout the book. The first chapter introduces some of the elementary concepts in neuroscience, including a whimsical analogy between TV sets and the retinal and cortical neural surfaces, as well as a description of the benefits of PDP (parallel distributed processing -- i.e., neural net computation). After an 'heroic-age-of-science' section, in which Churchland points out that seemingly controversial hypotheses have proven correct in the past and so might be expected to be proven correct in the future as well, the main objectives of the book are outlined. These are, generally speaking, introducing the general reader to the new theories and results from neurocomputational research (both natural and artificial), and, he hopes, motivating him or her to rethink our mental lives and to participate in the debates surrounding these new developments.

The second chapter deals with the important concept of vector coding, which is illustrated through examples such as the encoding of smells and faces. By not addressing the differences between vector coding and local coding, Paul Churchland has avoided a tricky question (for a discussion see da F. Costa, 1995). There is little doubt that vector coding (the representation of information in multi-dimensional state spaces) is an important computational and representational concept, as it has been known by researchers in statistic pattern recognition for a long time.

Chapter 3 addresses the issue of vector processing. It starts by presenting a very simple example of how a neuron can be configured to perform simple pattern recognition and follows this by discussing Cottrell's neural system for face recognition. The important concepts of synaptic adjustment (represented here by the backpropagation paradigm), internal coding and distributed representations, the emergence of categories, and inductive inference are also presented and illustrated in a didactic and well-motivated fashion. Two observations, however. The first relates to the fact that the network for face recognition has produced receptive fields (an interesting concept from neuroscience of the sensory region which activates a neuron, and which has not been introduced in EORSOS) that extend throughout the whole image area, thus enhancing fault-tolerance. By not emphasizing that such an emergent feature is to a large extent determined by the way in which the image space is covered by the first neural layer, Churchland seems to imply that such extensive receptive fields are common in neural systems. However, had a more localized mapping been used instead, more localized receptive fields would have been obtained which, in addition to being biologically more realistic, would also provide fault-tolerance and account for more effective (less redundant) and compact hierarchical representations of the visual information (e.g., in terms of straight line segments). Paul Churchland does observe that the neural structure for face recognition under consideration is unlikely to be faithful to nature -- however, ironically, only with respect to its third and fourth layers. The second observation regarding this chapter is that Cottrell's neural structure has been presented in EORSOS as a kind of major novel breakthrough (see the second paragraph in page 40). There is no doubt that Cottrell's is an interesting application, but it does not depart substantially from the traditional feedforward model. In addition, there are a number of alternative neural approaches to

face recognition capable of reasonable performance, such as those developed by I. Alexander and T. Poggio, which have been around for a long time.

Neural vector processing is further explored in Chapter 4, again from the exclusive perspective of feedforward neural networks, as a means of imitating parts of the brain. The first example addressed is Paul Churchland's model for stereo cortical processing. This model, which relies upon a specific geometry of synaptic connections (greatly simplifying the problem), is used to introduce the concepts of excitatory and inhibitory neural interconnections. Churchland's 'Lilliputian effect' illustrating stereoscopic vision (a pair of images of Manhattan island as seen by a giant) is remarkable and is itself worth the stereoscope which accompanies the book. The second example consists of Sejnowski and Gorman's network for sonar signature classification, a nice illustration of a situation where an artificial neural solution has beaten human expertise. NETtalk is taken for the third case study, whose principal merits reside in illustrating the potential of artificial neural nets for learning (although in supervised fashion) and in building categories. However, it should be noted that its initial problem (getting a computer to read printed text aloud) is in fact not as complex as it may at first appear, for this is a problem that could have been solved straightforwardly by the simple alternative of table look-up, i.e., by adopting a database containing the correct pronunciation for each possible English word. The fourth example, namely sensorimotor integration through vector transformation, is nicely illustrated with the aid of Paul Churchland's famous crab. Although the CalTech bias exhibited by the selection of examples does support a nice didactic illustration of several important aspects of neural nets, I felt that the inclusion of some alternative models -- such as Kohonen's self-organizing maps and the more recently introduced radial-basis neural networks -- would have served quite as well, while also introducing some of promising biological features by which they have been inspired.

Recurrent networks are presented in Chapter 5 mainly as a means of implementing the temporal dimension missing in the feedforward paradigm. The potential of such nets for generating and recognizing sequential patterns as well as their relationship with the ambiguous figure/background phenomenon are neatly described. This chapter also presents some exciting ideas relating recurrent networks to recognition, understanding and scientific progress, particularly from the perspective of vector coding and completion.

The title of Chapter 6, namely 'The Neural Representation of the Social World,' is self-explanatory. Here it is argued that recurrent networks as well as the related processing paradigms (e.g. vector processing, vector completion) are the principal mechanisms through which higher level activities such as language and social integration are achieved by humans. This chapter starts by briefly presenting and discussing Cottrell's EMPATH network, a variant of the face recognition system discussed above, that is dedicated to emotion recognition. Churchland uses EMPATH to illustrate a situation in which a neural network has proven to perform in a behavioral task (narrowly construed) almost as well as a human. However, the net is in fact doing nothing other than computing the spatial correlations between sets of pixels in images that are nicely registered (i.e., spatially aligned) and are devoid of any substantial geometrical transformations -- limitations

which human performance considerably outstrips. In addition to performing expert classification of emotions, and even of more sophisticated patterns, humans can precisely characterize such emotions and their contexts. It could be argued that a system for face recognition does not need to do anything more complex than to classify facial patterns by outputting a code number. Such a point of view, one that is in fact shared by a large part of the pattern recognition community, may itself be one of the reasons for the current lack of powerful and general recognition systems. For how can a hysterical laugh be distinguished of a common laugh if not by analyzing its whole context? Indeed, pattern recognition should not be limited to trivial object classification, but should also involve the perception of the multi-dimensional universe of related aspects, such as the function of the object, its typical context, its relationship with the rest of the world, its category, temporal dimension, and so on. Considering that such comprehensive knowledge is essential as a means of enhancing the recognition system performance, it may take some time before neural networks are developed which are fully able to perform in complex situations such as dealing with social representations.

The remainder of Chapter 6 deals with human linguistic ability, a discussion of Elman's neural approach to grammatical discrimination and the issue of moral perception and understanding. Making a long story short, this section is dedicated to the hypothesis that linguistic and moral skills are not rule-based, but are related to hierarchical learned prototypes stored in the neural system. The truth may, however, prove to lie somewhere in between rules and prototypes, for we do employ rules from time to time.

Chapter 7 treats many problems which can affect the human brain and behavior. This chapter starts by presenting in an uncomplicated manner some of the techniques which have been used as means of unveiling the anatomical and functional structure of the human brain, namely CAT, PET and MRI scans. Then, a number of diseases such as apraxia, MS, ALS, the striking cases of blindness denial and hemi-neglect, schizophrenia, manic-depression and depressive illnesses are characterized and discussed. It is stated that a better knowledge of biological neural networks can help us to develop improved treatments for such disorders, a prospect that no reasonable person would dare deny nowadays.

3. Seated Souls and Their Social and Moral Implications

The first chapter in the second part of EORSOS is fittingly entitled 'The Puzzle of Consciousness.' Not so fittingly, a variant of the 'heroic-age-of-science' argument -- in this case with respect to the fact that vitalism in biology has been replaced in surprising fashion by physical processes such as DNA -- is invoked as the means of introducing the controversies of consciousness research to the general reader. The great shortcoming of this approach, as expressed by Paul Churchland himself, is that such parallels provide no implications at all for the nature of consciousness. So, why to use them? In the following four sections of EORSOS, Leibnitz's mite example (imagine yourself as small as a mite wandering around inside your brain; presumably you will fail to find anything answering

to conscious experience), Nagel's bat, Jackson's neuroscientist Mary and Searle's Chinese room are all briefly described and dismissed (in different counterarguments). The inclusion of the relatively less known and simple Leibnitz mite argument is used as a kind of basic metaphor to undermine the subsequent philosophical positions. Although Paul Churchland's dismissal of such positions is reasonable from the perspective of purely mechanical aspects of consciousness, I feel that the key point of consciousness, namely its qualitative experiential aspect (e.g. Chalmers, 1993, 1996), has been almost completely missed. Having bypassed that tricky question, the subsequent sections in Churchland's book become constrained to the classical scheme identified by Chalmers (1996), i.e., to showing that the mechanical aspects of consciousness can be physically explained and even artificially implemented, in this specific case by using recurrent neural networks.

The issue whether an electronic machine can be conscious or not provides the theme for the next chapter, which starts with an interesting report regarding Paul Churchland's participation as an observer in the hidden room of Loebner's annual Turing Test Competition in 1993. The fact that five judges took one of the human participants to be a machine is understood by Paul Churchland as one of the indications that the Turing test should be reformulated in order to incorporate other cognitive features such as learning, recognition and conceptual change. Perhaps the cleverest feature of the Turing test is that it relies entirely on the fact that human intelligence is a relative concept and a concept that can best be assessed by humans. The next issue addressed in Chapter 9 consists of Mead's silicon retina, which is presented as a successful attempt to emulate a natural neural network, in this case the primate retina, in a dedicated and highly compact microchip. Some traditional objections to machine intelligence, namely how to provide machine states with meaning, the capacity for doing mathematics and the qualitative characters of conscious experience, are discussed subsequently. While the first two issues are represented as a critical discussion of Searle's and Penrose's positions respectively, Churchland approaches the third issue from the perspective of the 'out of fashion' functionalist approach to consciousness and, since according to Churchland this endeavor seems to have failed (cf. p. 251), the conclusion is (presumably) that the qualitative aspects, as well as other major features of human consciousness, are nothing else but a direct consequence of the physical configuration of the nervous system.

The next chapter -- 'Language, Science, Politics, and Art' -- ranges over all of the broad-brush themes. Strictly speaking, religion is also addressed. Issues such as intelligence differences between humans and other animal species, the question whether language is unique to humans and Dennett's language-oriented theory of consciousness are discussed in an enlightening fashion. The rest of the chapter addresses the hypothesis that scientific creativity, cognitive and moral abilities and even the arts can be explained from the perspective of neurocomputation.

The last chapter, entitled 'Neurotechnology and Human Life,' addresses the implications that the advances from neuroscience and neurocomputation may have on our lives, including their medical, legal and scientific impact as well as that upon our own self-concept. Churchland reasonably forecasts that applications of artificial recurrent networks

will include medical diagnosis and treatment as well as the assessment of criminals. The controversial problems of abortion, corrective policies, and scientific advancement are also discussed from the neurocomputational perspective. Although many of the positions expressed in this chapter are provocative, there is little doubt that advances from neurocomputational research, not to mention the progresses in genetic engineering, are bound to change the way in which we see the world and ourselves. Whether this might also lead to an improved moral and social standards is considerably less clear. Indeed, to judge from previous interference of science in such areas (e.g., phrenology, polygraphs, Bell curves and other freaks from the scientific menagerie), the prospects here may not be so rosy.

4. Concluding Remarks

Churchland's EORSOS is undeniably a well motivated and clearly written book. While its first part provides a nice introductory survey to neuroscience and neurocomputing, including the important concepts of vector coding, vector completion, PDP and recurrent networks, its second part covers the implications that such neurocomputing concepts the author believes will have on consciousness research, morality, science and society. Through the extensive use of graphics, metaphors, case studies and an excellent didactic technique Paul Churchland has successfully accomplished his main objectives in writing this exciting and provocative book. However, there are occasions throughout EORSOS in which artificial neural networks, and particularly those that goes under the name of recurrent networks, have been presented in a somewhat overenthusiastic fashion. Indeed, one of the positions defended in this book is that the whole plethora of remarkable behavioral and intellectual abilities exhibited by humans can be explained in terms of (artificial and natural) recurrent neural networks. This point seems to be supported by the following line of thought: (i) the recurrent nets in the human brain are responsible for virtually all of our intellectual abilities; (ii) recurrent networks are defined to be similar to those in the human brain; (iii) artificial recurrent networks are consequently capable of emulating the brain. The problem resides in hypothesis (ii), since there is an enormous gap between biological networks and the current artificial neural models, recurrent or not. (And, indeed, in connectionist research recurrent nets are not at all *defined* as being similar to the brain: recurrent nets simply add to artificial feedforward neural nets a feedback mechanism for retaining context, in a kind of 'short term memory'; cf. Ellman, 1990.) Consider the following facts:

(A) If the current neural paradigm is so good, why do we have no artificial neural structures capable of addressing actually challenging problems such as versatile pattern recognition and vision? Whereas it is relatively easy to design dedicated pattern recognition systems capable of 90% or even 98% correct classifications, it becomes exceedingly difficult to improve on such performance. It is likely that near 100% performance can only be achieved by incorporating additional information such as the spatio-temporal context (see above) of the objects under analysis, a task that is generally accomplished with expertise by human recurrent networks. This is not to say that artificial recurrent networks will never be able to perform such complex tasks, but only that the current nets do not.

(B) Most of the neural network techniques are not well understood in mathematical terms. The major problem is that we need such a knowledge in order to validate each neural network with respect to specific applications, particularly the most critical ones (e.g., criminal analysis and medical care), since artificial neural nets are plagued by shortcomings such as the plasticity/stability compromise, inherent noise, and many other problems. It is true that our recurrent nets are also not completely reliable, but we need at least to determine that the error rates of such artificial neural systems are not worse than ours.

(C) There has been a general lack of effort towards assessing the performance of artificial neural networks comparatively with alternative approaches to pattern recognition and classification, and Paul Churchland provides no exception to the trend. The fact of the matter is that alternative processing strategies such as statistical and probabilistic classification (in, for example, the computer programs: AUTOCLASS, Cheeseman et al., 1988; C4.5, Quinlan, 1993; and SNOB, Wallace, 1990; finally see Michie et al., 1994, for an experimental comparison of statistical, AI and neural network techniques), an area which is itself a subject of intense research, do provide effective solutions to many, if not all, of the problems to which artificial neural networks have been applied. Is it sensible to overlook such a great asset? It is true that statistical approaches are not inherently parallel, but the area of concurrent processing (e.g. pipelining, systolic arrays, vector processing) and fault-tolerant systems seems to have come of age (the ubiquitous Intel processors used in personal computers present in themselves a definite trend toward becoming truly parallel systems). It is a pity that statistical artificial intelligence has been overlooked in EORSOS, for there is a strong tendency towards hybrid systems in which the advantages and disadvantages of the neural and statistical approaches could be used complementary fashion.

In brief, although there is little doubt that biological recurrent networks are capable of producing human abilities, this endeavor is beyond any doubt far beyond the capabilities of the currently available artificial recurrent neural networks.

As far as consciousness is concerned, I feel that its most tricky feature, namely qualia or qualitative experience, has been overlooked, perhaps due to the same reasoning as that outlined above: (i) recurrent networks in the human brain are capable of sensation; (ii) artificial recurrent networks are defined to be analogous to those in the human brain; (iii) artificial recurrent networks are consequently capable of sensation. While this is possibly true as far as the mechanical aspects of consciousness are concerned, such a reasoning appears to rely on the very functionalism which Churchland elsewhere rejects, since it presupposes that all aspects of consciousness depend upon the functional features of the brain, which can be recapitulated in the hardware of artificial recurrent neural networks. The 'hard problem' of Chalmers (1996), of spanning the gap between function and feel, is hardly even addressed. But it is precisely here that the controversies about consciousness, swirling around Searle, Penrose and others, have found their energy.

I conclude that, in spite of all the well-intended approaches to the problems of

consciousness such as Churchland's here, the qualitative aspect of consciousness remains as puzzling as ever. Despite all the scientific-technological advances over the centuries, perhaps we have reached a point of being forced to simply acknowledge our ignorance without prospect of any quick improvement in that state. This is by no means accepting 'mysterianism' or supernatural concepts. On the contrary, it is only by precisely characterizing our lack of knowledge that solid basis for further scientific advances can be achieved. After all, there are worse things than not knowing the answers to our questions.

Acknowledgements

The author is indebted to FAPESP and CNPq for financial help and to Dr K. Korb (Monash University, Australia) for invaluable suggestions and enhancements to the present text.

References

Chalmers, D. J. (1993). *Toward a theory of consciousness.* PhD thesis, Indiana University.

Chalmers, D. J. (1996). 'Facing up to the problem of consciousness.' In *Proceedings: Toward a scientific basis for consciousness.* MIT Press.

Cheeseman, P., Self, M., Stutz, J., Kelly, J., Taylor, W. and D. Freeman (1988) 'Autoclass: a Bayesian classification system,' *5th Int. Conf. on Machine Learning.*

Churchland, P.M. (1995). *The engine of reason, the seat of the soul.* MIT Press.

da F. Costa, L. (1995). 'On the most efficient and versatile real-time image processing system - A review of *The computational brain*,' *Real-Time Imaging, 1,* 87-90.

Ellman, J. (1990) 'Finding structure in time,' *Cognitive Science, 14,* 179-211.

Michie, D., Spieghalter,D.J. and Taylor, C.C. (eds.) (1994) *Machine learning, neural and statistical classification.* Ellis Horwood.

Quinlan, J.R. (1993) *C4.5: programs for machine learning.* San Mateo, Calif.: Morgan Kaufmann.

Searle, J. R. (1992). *The rediscovery of the mind.* MIT Press.

Wallace, C.S. (1990) 'Classification by minimum-message-length encoding,' *Advances in Computing and Information - ICCI '90* (pp. 72-81). Springer Verlag.