

# The Engine of Reason, the Seat of the Soul

by Paul Churchland

Reviewed by Robert M. French

(Review of Paul M. Churchland, *The Engine of Reason, the Seat of the Soul* The MIT Press, Cambridge, MA. In *Minds & Machines*, 6(3), 1996. pp. 416-421.)

For the uninitiated, there are two major tendencies in the modeling of human cognition. The older, traditional school believes, in essence, that full human cognition can be modeled by dividing the world up into distinct entities -- called *symbols*-- such as ‘dog’, ‘cat’, ‘run’, ‘bite’, ‘happy’, ‘tumbleweed’, and so on, and then manipulating this vast set of symbols by a very complex and very subtle set of rules. The opposing school claims that this system, while it might be good at concluding that Paris is the capital of France or that there must be blood flowing in the left-rear leg of a cow, can never capture the full measure -- indeed, the essence -- of human cognition. For them, the essential features of cognition emerge from the combined effects of myriad, tiny actions far below the surface of consciousness. This is the camp to which Paul Churchland belongs.

Now, let us turn to Churchland’s book, *The Engine of Reason, the Seat of the Soul*. It is a clearly written, easily understood presentation of some of the most important ideas and impressive contributions of connectionism. He leads the reader step by step through various kinds of ‘connectionist’ networks, from the simple backpropagation networks developed in the early 1980’s through the recurrent networks that were developed in response to problems that the simpler networks could not handle. He extrapolates from these networks to vastly larger, vastly more powerful networks that he believes will ultimately lead to a full simulation of human cognition. He describes a number of fascinating case studies, including Charles Rosenberg and Terry Sejnowski’s NETtalk, an early connectionist network that learned to pronounce English words. His excellent discussion of NETtalk accurately captures the excitement that this seminal program generated around 1986 when it first forced many traditional artificial intelligence researchers to sit up and take connectionism seriously. Perhaps more than any other program in the field, NETtalk was responsible for the tremendous surge of interest in connectionism and in emergent (‘bottom up’) models of cognition.

The book includes a detailed and extremely interesting chapter on connectionist approaches to stereoscopic vision, detection of mines by submarines, pronunciation and, even, crab movement! Churchland carefully explains why recurrent networks, as opposed to simple backpropagation networks, must be used to process sequences of events. There are chapters on brain dysfunction, consciousness (including some ground-breaking work by Rodolfo Llinas on neo-cortical oscillations and the Crick-Koch hypothesis that these oscillations may be the seat of consciousness), and potential technical uses of neural networks, including medical diagnosis. It all makes for truly fascinating reading.

There are, however, a number of important problems with this book that cannot be ignored.

To begin with, the book all too frequently reads like an ‘infomercial’ for connectionism and ‘prototype vectors’. Infomercials, as everyone knows, contain a certain amount of truth wrapped in hyperbole and sold with evangelistic fervor. This is emphatically not what the neural network research program needs. When enthusiasm for an idea causes its proponents to intentionally downplay, overlook or conceal major difficulties with it, the inevitable result is not only bad science, but a disillusioned public. The first page of the book is almost certainly the worst of all. Churchland writes:

“... we are now in a position to explain how our vivid sensory experience arises in the sensory cortex of our brains: how the smell of baking bread, the sound of an oboe, the taste of a peach, and the color of a sunrise are all embodied in a vast chorus of neural

activity. We now have the resources to explain how the motor cortex, the cerebellum, and the spinal cord conduct an orchestra of muscles to perform the cheetah's dash, the falcon's strike, or the ballerina's dying swan. More centrally, we can now understand how the infant brain slowly develops a framework of concepts with which to comprehend the world. And we can see how the matured brain deploys that framework almost instantaneously: to recognize similarities, to grasp analogies, and to anticipate both the immediate and the distant future."

Poetic, perhaps, but absolutely false. And what's more, Churchland is far too well informed not to be aware that it is false. There is, of course, a trivial sense in which it is true. Scientists can say, with complete, but trivial accuracy: "We *are* in a position to explain all these wonderful things because we know their cause -- namely, the interaction of neurons in the brain." Or, if they want to sound more scientific (and obfuscatory), they can say, "Sequences of transformations on vectors of length  $10^{14}$  make them happen." Well, that's correct, too, but still not very helpful. And, moreover, this is surely not what Churchland means. He means, I presume, that science is actually within striking range of understanding the ballerina's dying swan, human analogy-making, and the smell of baking bread. And that is simply absurd. We are *nowhere near* understanding these things and, furthermore, it is irresponsible to claim otherwise to a readership desirous of a better idea of current scientific progress in understanding human cognition.

In *The Engine of Reason, the Seat of the Soul*, Churchland systematically sidesteps any of the really hard problems that face researchers modeling cognition, including one of the biggest problems of all, the problem of representation. As Churchland says, connectionist models do an excellent job in the presence of degraded input -- something that traditional symbolic models did very poorly -- and they do, indeed, categorize and generalize very nicely in many cases. But this is a far cry from what humans do. He discusses "inductive inference, network style" and talks about recognizing an object if 20% of its information has been removed. That's a splendid achievement, but humans can do not only far better (spotting a mere 2% of my keys on my cluttered desk, is usually sufficient for me to recognize and retrieve them) but also far worse (I have stared straight at someone I knew well -- i.e., 100% of the information was available at the retina -- but did not recognize him because I absolutely did not expect to see him at that time and place).

In Chapter 5, Churchland suggests that the birth of a new scientific theory -- for example, Einstein's non-Euclidean spacetime interpretation of the cosmos -- is, in essence, a matter of recognizing "some unfamiliar, puzzling, or otherwise problematic situation as being an instance or example of something well known." Up to this point, we are basically in agreement. We part company, however, when Churchland claims that this recognition is merely a matter of "... the brain's *vector completion* of partial or degraded inputs...". His key idea is that great leaps of creative understanding are due to activation flowing around in the immense recurrent network which is our brain until it "finally activates some vector close to one of its antecedently learned prototypes."

Great insights are, indeed, due to great analogies and great analogies arise from representing a complex situation in a *novel* manner and putting that novel representation into correspondence with "an instance or example of something well known." We understand the world by continually emphasizing different aspects of our long-term memory representations. Pick up any ordinary object, say, a credit card and consider the following sentence: "A credit card is like an X," where X can be any object of your choice. And then observe how you focus on different, often novel, unsuspected aspects of the representation of "credit card." Some examples: A credit card is like a check book. A credit card is like a door key. A credit card is like a hospital. A credit card is like a ruler. A credit card is like a doormat. A credit card is like a rose. A credit card is like a banana peel. A credit card is like a switch-blade knife... The list is endless, any noun can replace X, but you will always be able to transfer some facet of your long-term memory representation of "credit card" (perhaps very

stretched, but that is beside the point) to working memory in order to be able to say why a credit card is like that object.

But in Churchland's world of recurrent networks and prototype vectors there is no such distinction between passive long-term memory (LTM) representations -- representations that conceivably include your entire life experience -- and working memory (WM) where the context-appropriate subsets of those long-term-memory representations are activated and used. Understanding how this LTM-to-WM transfer works is one of the big questions of cognitive modeling and there is not even a hint of it in Churchland's book. In Churchland view, there are only long-term memory representations -- "prototype vectors." Churchland would presumably argue that this distinction is achieved when the brain activates certain parts of its prototype vectors and not others, depending on the input context. True, of course, but how in heaven's name could a connectionist network be organized so as to achieve this? Churchland doesn't not even touch on this major issue.

One has the impression in reading Churchland that as soon as the brain notices enough overlap between, say, the prototype vector for "four-dimensional non-Euclidean geometry with three spatial dimensions and one temporal dimension" and the prototype vector for "the universe", out will pop Einstein's new vision of the cosmos. Dubious. But why? Because either each prototype vector must consist of the entire state of the brain -- in which case, there are at least  $2^{10^{14}}$  prototype vectors and the idea that activation will just spread to the right vector, thereby leading to the discovery of the General Relativity, is ludicrous -- or prototype vectors are smaller than that, only incorporating the "essential features" of a situations (as the name "prototype" would imply). But in the latter case, Churchland is brought face to face with the same problem that plagued traditional artificial intelligence -- namely, the impossibility of objectively determining features of an object that characterize it in a context-independent manner.

Let's bring this problem down from Einstein's revolutionary view of the universe to the most ordinary of utterances: "After the Christmas holidays my bathroom scale is my worst enemy," We all know exactly what this sentence means. But what a priori representations of "bathroom scale" and "worst enemy" could allow us to understand this simple expression? It would have to include knowledge about the tradition of big meals and excessive eating at Christmas, about people's concerns about being overweight, about irony, as well as subtle and complex knowledge about battles, enemies and competition in order to make sense of the idea of a hostile encounter between you and your bathroom scale, etc. ALL of this must be included in the prototype vectors for "bathroom scale" and for "worst enemy" in order to understand this simple utterance. And how are just the right parts of each of these representations selected and put into correspondence in order for us to (instantly) understand the sentence?

I am not, of course, saying that recurrent networks could never do this. They already can -- you just did it without a second's hesitation -- but the question is *how* do they do it? It is not enough to enthusiastically wave the magic wand of "vector completion."

And there are other important problems that Churchland avoids mentioning. For example, the very thing that gives connectionist networks their power to generalize, to handle degraded input, and so on, also causes catastrophic interference. In other words, if a network learns to associate a hundred symptoms with a hundred diseases, adding a few new symptoms associated with a few new diseases can cause the network to *completely and suddenly* forget all one hundred of the previously learned symptom-disease associations. From both a psychological and practical standpoint, this is decidedly poor design. Certain authors have suggested that the modular separation of the hippocampus and the neo-cortex is the brain's way of dealing with this problem -- but we see no mention in Churchland's book that any such problem, known and studied by connectionists since 1989, even exists.

Another "oversight" in Churchland's discussion of neural networks is episodic memory. The learning algorithms of the networks that Churchland describes work by gradually modifying the strengths of the connections between units. But these algorithms are very poorly suited to the

problem of one-trial learning or episodic memory. How, in terms of standard, connectionist learning procedures, is it possible for me to remember, with vivid and absolute clarity, the image of a car rammed beneath a semi-trailer, a scene that I observed for no more than ten seconds twelve years ago?

Simply put, Churchland puts a spin on the facts of neural network research that is, at times, very misleading. One can be enthusiastic about connectionism -- and there is every reason to be -- without falling into the trap of infomercial-like hype, of which there is far too much in this book. One would have thought that the failed promises of the early proponents of artificial intelligence and of the earliest "connectionist," Frank Rosenblatt, who invented the forerunner of today's neural networks, would have served as a warning to those who would begin yet another round of propaganda. Research in modeling cognition has a long and difficult road ahead of it and not only the triumphs, but also the difficulties must be discussed if we are to progress. To be fair, Churchland does occasionally acknowledge difficulties when he is able to rein in his exuberance. For example, he concludes Chapter 9 with a far more sober statement than the one on the first page of his book: "Could an electronic machine be conscious? It rather looks that way. Will it happen soon? Probably not, although small steps will continue to be taken." But this type of cautious language is all too infrequent in *The Engine of Reason, the Seat of the Soul*.

In short, Paul Churchland has written a fascinating book for the lay audience describing some of the major triumphs of neural-network modeling. But all too often he has let his enthusiasm get the best of him and hardly ever mentions the really hard problems facing researchers in the field of neural modeling. He has a long way to go before he can claim to have made the case that neural networks hold the answers to Everything Concerning Cognition.