From paleo to neo connectionism,
reprinted from
In G. van der Vijver, ed., *New Perspectives on Cybernetics,* Kluwer, Dordrecht, 1992, p. 125-146.

FROM PALEO TO NEO-CONNECTIONISM*

The very label 'connectionism' points to a simple question, whether or not a prefixed 'neo' emphasizes it: what are the links, and oppositions, between the old and the new connectionism—between say McCulloch & Pitts and Hopfield, or Rosenblatt and PDP[1]? This paper will not be an

---

* The present paper grew out of talks given in 1987 in Jerusalem (Mishkenot Sha'ahanim workshop on "Minds and machines", April), Cerisy-la-Salle (conference on "Approches de la cognition", June) and Gent (conference on "Communication and cognition", December).

[1] The word 'connectionist' is used by Donald Hebb in 1949 and by Frank Rosenblatt in 1958, as James Anderson and Edward Rosenfeld, the editors of *Neurocomputing,* point out. Since I delivered a talk under the title of the present paper at the Gent conference, MIT Press published this most precious anthology (hereinafter A & R), and thus made the job of tracing the roots and paths of neural network research at once considerably easier, by providing, together with knowledgeable introductions, all the major papers, and quite a bit more difficult, by ruling out any simplistic history one might have been tempted to propose. One simple idea I have been able to retain is that of a three-stage development : the first period, fairly continuous, slow but extremely fertile, extends from the 1943 paper of McCulloch and Pitts's to Minsky and Papert's 1969 book *Perceptrons;* it is followed by a transitory decade leading in turn to the present explosive stage (heralded by a number of bangs -- the Hinton & Anderson collection in 1979, Grossberg's 1980 paper, Hopfield's 1982 paper, Feldman's and Ballard's 1982 survey). In this paper, 'paleoconnectionism' will refer to the first period, especially to the fifteen years 1943-1958, and 'neoconnectionism' roughly to the last twelve years, 1979-present. An important and difficult question is raised precisely by the transition and by the role played by investigators who didn't stop being 'connectionists' in 1969 nor waited until 1982 to become members. This question will not be addressed here, but it is likely that a full treatment of the questions which I do attempt to address would benefit from an adequate account of that *interregnum.*

exercise in "Contrast and Compare", nor will it attempt to list and relate all the main ideas and trends as they have historically emerged. Instead, it will focus on a small number of puzzles, such as the following, expressed in a deliberately provocative, and somewhat biased way:

— How can one seriously ask today whether nets can *really* do logic, while as everyone knows nets in McCulloch's days did exactly that (and only that)?

— More generally, how is it that the problem(s) of cognition is (are) nowhere near solved half a century after they have been proclaimed solved "in principle" by McCulloch (and von Neumann), while neither of the main contenders to cognitive truth today—classical computationalism and connectionism—have unequivocally disowned McCulloch?

— How could the mere adjunction of some "hidden" units between the layers of the perceptron (actually envisaged by its alleged grave-diggers, Minsky and Papert), coupled with faster (von Neumann) computers, change the face of connectionism, of AI, and perhaps even of cognitive science at large[2]?

This paper therefore is not an attempt at serious intellectual history. It is an attempt at clarifying some issues in the foundations of contemporary cognitive science by examining certain aspects of earlier conceptual frameworks for the study of the "mind/brain". No doubt the latter enterprise would benefit greatly from progress in the former:

---

[2] No attempt will be made in this paper to sketch its general framework, *viz*. cognitive science, classical computationalism and connectionism. The references are too numerous to list; for the classical perspective, see for example Haugeland 1978, Fodor 1981, Pylyshyn 1984, for (neo)connectionism, see Rumelhart & McClelland 1986, Smolensky 1988, Amit 1989; for comparative studies, see Clark 1989, Andler 1986, Andler 1990 (mostly sympathetic to connectionism), and Fodor & Pylyshyn 1988 and the two other papers in Pinker & Mehler 1988 (mostly hostile to connectionism).

the present paper will surely require revisions in the light of genuine historiography[3].

1. TWO PERSPECTIVES ON MCCULLOCH AND PITTS'S NEURAL NETS

The harvest brought in by the great logicians of the 1930's was bountiful indeed. It included the notion of a formal system, of which the propositional calculus and the predicate calculus became the most important examples, and the notion of a recursive function, which provides the basis for a universal, completely general theory of computation. There are important links between the two notions, as was recognized from the beginning and has become ever more evident in the last twenty-five years or so. However, the basic fact remains that they are quite distinct. Not only are they *prima facie* as removed from one another as, say, the English language and the algorithm for long division, but it takes quite a bit of work to bring one to bear on the other. For example, to a large extent, Gödel's *tour de force* in his proof of the incompleteness theorem consisted in exploiting the computable nature of the morpho-syntactic operations in a formal system and the representability of recursive functions in any sufficiently strong formal arithmetic. This is one indication, among many, which should help the layman realize the distance separating the "and" in such textbook titles as *Computability and Logic*[4] from any version, however weak, of the "equal" sign.

In their famous 1943 paper, McCulloch and Pitts defend two claims. The first and presumably central one is indicated, if somewhat obscurely, by the title: "A logical calculus of the ideas immanent in nervous activity". The claim is that some neural nets "do" logic, more precisely, propositional logic. The second claim, which occurs at the

---

[3] Jean Mosconi's unpublished doctoral thesis (Université Paris I, 1988) is certainly the place to start. I deeply regret not having had a chance to read it before setting on this investigation.

[4] E.g. Boolos & Jeffrey 1974.

very end of the paper and comes, by way of proof, with a mere "It is easily shown that…", is that a neural net can (more or less[5]) compute any number that a Turing machine can compute.

The resulting trace in collective memory is thus ambivalent: close one eye, and you see "embodied" logic; close the other, "embodied" computation appears. It is somewhat of a mystery to the present writer that, of the several commentators he has read (first and foremost the authors of the foreword to the new edition—Jerome Lettvin—and of the introduction to the first edition—Seymour Papert—of McCulloch's collected papers, and the editors of *Neurocomputing)*, none has bothered to peel apart the two threads. Perhaps the distinction has been seen as too obvious to deserve attention—as indeed it is to anyone who examines the paper from the standpoint of logic.

However, one may suspect that some deeper confusion—or disagreement—is lurking, for example when one recalls one of Jerry Fodor's favorite mottoes: "No representation, no computation"[6]. Stripped from context, it seems to imply that representation goes hand in hand with computation. If you think of what is represented as, say, states of affairs, or propositions, then you are tempted to confuse calculus on propositions with general computation, and then perhaps just logic with computation. Then you are bound to equate "embodied" logic with "embodied" computation, and see *one* result in the McCulloch-Pitts paper where you should see two admittedly related, but conceptually distinct, results. And to confuse two very different notions of completeness or universality: one which applies to a set of connectives, such as {**not, and, or**}, and indicates the ability (of that set, or of any mechanism, say a neural net, which "realizes" it) to express any complex

---

[5] Depending on whether it does or doesn't have circles, and is or isn't outfitted with a tape and the functional equivalent of a read-write head.

[6] See *e.g.* Fodor 1981, Fodor 1985, Fodor & Pylyshyn 1988.

proposition, the other which applies to a computing machine and indicates the ability (of that machine, or of any mechanism, say a neural net, which emulates it) to compute *any* function which can be computed, that is to provide the value $f(x)$ of any given computable function f on any given integer x belonging to the domain of f.

Tracing the root of this confusion will help us grasp what is perhaps the most important difference between the old and new connectionisms. So let us reexamine briefly the two theses put forward in the 1943 paper.

a) McCulloch-Pitts nets as logic machines

Take a neural net comprising at least three neurons A, B, C, and assume that A's being active means that **p** (a proposition), B's being active means that **q** (some other proposition), and that C is active except when A is active and B isn't. Clearly, then, that net, under the proper description, "embodies" the logical relation "$\mathbf{p} \rightarrow \mathbf{q}$". Assume further that the brain comprises a structure which, for some purposes, is faithfully modelled by that neural net. Then you have an explanation of the brain's ability to effectuate logical implication; *ergo,* the *mind*'s ability to grasp the relation of logical implication between propositions is accounted for. Similarly, of course, for any other logical relation one might think of. "Thus the […] formal aspect of that activity which we are wont to call *mental* [is] rigorously deducible from present neurophysiology."[7]

The contemporary reader, duly warned by McCulloch himself ("Don't bite my finger, look where I'm pointing!"[8]), might as well pause at this point to catch his breath. For just about everything seems to be wrong with this line of reasoning. First, logic isn't just propositional calculus. Second, what the above net instantiates, at best, is not logical

---

[7] McCulloch & Pitts 1943, in McCulloch 1965/1988: 38.

[8] Quoted by S. Papert, in McCulloch 1965/1988: xxviii.

implication in general, but the complex formula **p** → **q**, with **p** and **q** fixed (they are the "content" of the activity of neurons A and B respectively). Third, in order to be a plausible candidate for the job of "embodying" **p** → **q**, the net should somehow, somewhere detain an intrinsic property making it specifically the embodiment of **p** → **q**, rather than that of any other complex formula built from **p** and **q**, such as the (truth-functionally) equivalent ¬**p** ∨ **q** or perhaps even the non-equivalent **p** ∨ **q**. For **p** → **q** itself, whether regarded as a platonic proposition or as a sequence of symbols, is intrinsically, *non-relationally* distinct from any morphologically distinct proposition or formula. But what property could do the job for our net is, to say the least, unobvious[9].

The first two objections can be disposed with fairly quickly in this context. As for the first, it should be admitted that logic, ever since cognitive science came into being, has been reduced to propositional calculus (be it somewhat enriched, *e.g.* modal, or deviant, *e.g.* probabilistic, intuitionistic etc.)— it would therefore be highly unfair to demand more of the pioneers than we generally demand of ourselves. Besides, McCulloch and Pitts had not lost awareness of the need for quantification; as pointed out by McCulloch[10], nets with circles can give rise to reverberation, which is a form or model of memory, which can in turn be seen as providing a temporal equivalent to existential

[9] This point is made by Fodor 1986 (and again in Fodor & Pylyshyn 1988), but is meant by him to apply (with supposedly devastating effects) to the nets of *neo*connectionism. This struck me, when I first came upon it, as a complete misunderstanding of *that* kind of connectionism, and led me to consider the problem of locating the difference which would explain why a perfectly cogent argument against nets as construed by McCulloch and Pitts appears out of place in the context of contemporary neural net research.

[10] McCulloch 1961, in McCulloch 1965/1988: 10.

quantification. This may be a bit swift, but allows us to at least suspend the objection.

The second concern could perhaps be alleviated as follows: after all, **p** and **q** cannot plausibly be thought of as completely fixed, else the net would be pretty useless: they presumably vary with time and circumstances (or at least their truth-values do, perhaps because they contain, be it implicitly, some indexical element: **p** might be "It is raining" or "It is raining here now"). But then why not let the contents of A and B vary completely freely? Then the net would "embody" $\mathbf{p} \rightarrow \mathbf{q}$ for *variable* **p** and **q** after all. Of course, this suggestion leads to the problem of distinguishing between the *activity* of neuron A and its *content,* which could be avoided as long as **p** was kept *perfectly* fixed—the activity of A *meant* **p** (or meant that the state of affairs expressed by **p** was "believed" to hold by the net, or by the brain).

But this, in another guise, is our third objection. In contemporary cognitive science, we would say that A *represents* **p** and that therefore A being active means that **p** is deemed true by the system. We thus distinguish between the represented proposition and its assumed truth value, and ask first in virtue of what A represents **p** rather than **p′** (and also why A represents anything at all), and only second in virtue of what A is active when, and only when, **p** is in fact true. In the case of neuron C, we would want to be in a position to assert that it represents $\mathbf{p} \rightarrow \mathbf{q}$, and we might then want to explain why C is on precisely when, in virtue of the states of neurons A and B, $\mathbf{p} \rightarrow \mathbf{q}$ is true, by the inner workings of the net ABC.

Now the question is whether there is anything in McCulloch and Pitts's avowed intentions, or in the structure of the nets themselves, that would allow us to make a similar move. In other words, is there some kind of representation intented or inscribed in the McCulloch-Pitts net? With this question in mind, we find that there are two readings of the 1943 paper, which we now examine in turn. There is a completely harmless way of construing the "immanent

calculus": propositional logic *models* cerebral activity. Thus, in their introduction, the authors write: "The «all-or-none» law of nervous activity is sufficient to insure that the activity of any neuron may be represented as a proposition. Physiological relations existing among nervous activities correspond, of course, to relations among the propositions; and the utility of the representation depends upon the identity of these relations with those of the logic of propositions."[11] In other words, McCulloch and Pitts do for neural nets—and therefore, or so they hope, for the brain (under description)— what Claude Shannon had done for electric circuits in his master's thesis[12] a few years earlier: use logic (actually, Boolean algebra) to describe or model the behavior of a complex system. The success of the approach, in either case, is based on the empirical fact that the physical magnitudes which characterize the system are "disciplined" by the *mathematical* laws or rules of propositional logic[13].

Such a reading is harmless conceptually, quite precious, as it will turn out, in the development of the digital computer concept, and completely uninformative from the cognitive or informational standpoint: neurons, or neuronal activities, on this reading, do not represent. Instead, they *are represented* by propositions and their truth values—just as parts of switching circuits can be, and many other things in the physical world.

---

[11] McCulloch 1965/1988: 21. The idea expressed in this quote is explicitly attributed to "one of [the authors]", *viz.* of course McCulloch (in McCulloch 1961, he tells the story of his discovery: "In 1929 it dawned on me...". McCulloch 1965/1988: 8-9). McCulloch was born in 1898, Pitts in 1923. In the conclusion of the paper (p.38) it is again asserted that "in psychology [...], the fundamental relations are those of two-valued logic."

[12] Shannon 1938.

[13] On the idea, crucial to modern science in general (and not just cognitive science), of the common "discipline" imposed by mathematics upon symbols and by nature upon the things symbolized, see Cummins 1989: 27-29.

We will return to this perspective, but first another, more informative and less harmless reading should be envisaged. Sometimes McCulloch and Pitts seem to suggest that the «all-or-none» character of neuronal activity makes it "inherently propositional", thus conferring upon "all psychic events […] an intentional, or «semiotic» character"[14]. It then sounds as if neurons were actually in the business of standing for states of affairs. Taken in isolation, out of context, such an idea would seem preposterous: the fact that sails come in white and non-white doesn't make them *inherently* propositional—it took a special convention between a father and a son to lead the former to believe, falsely as it turned out, that the blackness of a certain sail meant the demise of the latter. But Lettvin, in his foreword, provides essential background: "Once it was realized in the nineteenth century that nerve fibers conduct electric pulses and that the pulse trains on these fibers carry meaningful messages, the problem was to account for how such information was processed by the brain as nerve net."[15]

Thus neurons do after all represent, or process "meaningful messages". On that reading, there is, in the background, some theory of representation. But not only is it not provided; nowhere do McCulloch and Pitts show any sensitivity to the need to distinguish between the formal object **p**, or the "intention", the *"Sinn"* on one hand, and the truth value or *"Bedeutung"* of **p**[16] on the other**.** There lies, as we will see, a major difference with neo-connectionism. But right now it leads us to ask in what sense then the first cyberneticians took neural nets to "do" (propositional) logic.

---

[14] McCulloch 1965/1988: 37.

[15] McCulloch 1965/1988: viii.

[16] This, I presume, is the fault that Heinz von Foerster, the leader of the "second cybernetics", who founded and headed the Biological Computer Laboratory at the University of Illinois in the 60's, finds in McCulloch's concept of information: he confused it, or so claims von Foerster, with a mere signal. See Lévy 1985, Livet 1985, Livet 1990.

Althought the answer is quite obvious, let us take the historical route to (re)discover it. Lettvin provides yet another interpretation of what McCulloch and Pitts regarded the neuron as representing: he writes that "Warren and Walter were led [mostly by David Lloyd's work in 1939-41] to conceive of single neurons as […] acting as gates"[17]. But can a neuron represent both a proposition and a connective? Perhaps neurons such as A and B above represent propositions, while C represents material implication (→)? The answer is subtler. Let us state it first in approximate form, for the sake of clarity: a neuron *is* (or *embodies)* a connective, and its activity *represents* the result of applying the connective to the propositions which are fed to it—in other words, a neuron N embodies a certain function ***f*** and its being on or off represents the truth or falsity of the proposition ***f*** (**p**, **q**), where **p** and **q** are the propositions asserted by the activities of the neurons affering onto N.

Well, this is still not quite right, because ***f*** does not operate on *formulas,* but on 0's and 1's. So neurons embody not connectives, but Boolean functions, and operate not on formulas, but (at best) on truth values of formulas. Final correction: neurons being modelled (here) as threshold automata are not in themselves Boolean functions; these are realized by *nets* of formal neurons.

Thus we have finally reached the proper description of McCulloch-Pitts nets as logic machines: they embody, or realize, or instantiate Boolean functions. They should really be called Boolean machines. However, they can be regarded as performing "blind" or "animal" logic, provided an interpretation is somehow given to certain neurons, the "peripheral" ones which receive "information" from external sources; they can then be construed as *applied* Boolean machines.

---

[17] *Ibid.* It should be noted however that the 1943 paper contains no reference to Lloyd's work.

What is missing however from this model of nervous activity in order to constitute a neurophysiological account of genuine logical thinking, or cogitation, is an internal system of representations—an inner blackboard ready to receive the material inscriptions of the formulas whose truth values are being evaluated—and possibly others, kept in storage for future needs, for counterfactual reasoning, etc.

Incidentally, this model is not even a *plausible* neurophysiological account of blind logic, due to the random and unreliable nature of connectivity and single-neuron activity in the brain, as pointed out, in harsh terms, by second-generation connectionists such as Frank Rosenblatt[18]. Neo-connectionists will conclude that logical thinking does not occur at the aggregation or complexity level of single neurons, but (if it does occur at all) at a much higher level.

b) McCulloch-Pitts nets as approximations to Turing machines

All told, it is the above perspective on McCulloch-Pitts nets which seems to suffer from a rather extreme lack of plausibility. In fact, despite the very title of the 1943 paper, I used to think of it as plain slander, and only reluctantly admitted that McCulloch at least did sometimes believe that he and Pitts had shown how brains, thus minds, do (propositional) logic.

If one looks at the technical setting of the paper—a search for a description of the behavior of nets in terms of complex (temporal quantified) logical formulas, and conversely for nets whose behavior fits a given such formula—one gains a rather different outlook on their achievement. That this is the one which considered scientific judgment commands is confirmed by comments made some time later by such thinkers as von Neumann, Papert, and in fact McCulloch himself. Let us begin with the co-author of the famous paper:

---

[18] Rosenblatt 1958; in A & R: 93, col. 2.

"What Pitts and I had shown was that neurons that could be excited or inhibited, given a proper net, could extract any configuration of signals in its input. Because the form of the entire argument was strictly logical, and because Gödel had arithmetized logic, we had proved, in substance, the equivalence of all general Turing machines—man-made or begotten."[19] Although I would rather not be called upon to fully explicate this statement, its conclusion is clear enough: the "logical importance" of the 1943 paper is that the brain (or some subsystems of it) can be regarded as a Turing machine—in other words, as an "embodied" computer, and possibly as a universal one, *i.e.* one capable of computing any (computable) function whatsoever.

Papert, in his preface to the collected papers (published under McCulloch's own supervision), at once confirms the primacy of this perspective, introduces an important proviso and points to the (otherwise evident) link between the two perspectives: "When we reach the end of the paper, we are rewarded for the effort […] by seeing the first birth of a true mathematical idea: Between the class of trivial combinational functions computable by simple Boolean logic and the too general class of functions computable by Turing machines, there are intermediate classes of computability determined by the most universal and natural mathematical feature of the net—its finiteness."[20] Let us disregard the proviso (as devoid of relevance for our present purpose) and simply stress that once it is recognized that the net as logic machine really is a Boolean machine or computer, it takes but one small conceptual (though technically steep) step to the idea of the net as a (quasi) Turing machine.

But with von Neumann we now come to the heart of the problem posed by this sounder, mathematically richer outlook; in his address to the Hixon Symposium in 1948, he says: "[Their result] is that anything that can be exhaustively

---

[19] McCulloch 1961, in McCulloch 1965/1988: 9-10.

[20] McCulloch 1965/1988: xxvi.

and unambiguously described, anything that can be completely and unambiguously put into words, is *ipso facto* realizable by a suitable finite neural network."[21] How are we to understand the "-thing" in "anything"? Obviously as a *process.* We are provided, according to von Neumann, with a neurophysiological (or maybe just "neural") account of all mental processes. However, Turing machines operate on integers (or perhaps on marks), while mental processes operate on practically any domain and actually rather seldom on integers (and never on marks, except in tic-tac-toe!). Well, that can easily be fixed: just code the objects of the domain—with the help of some Gödel numbering, for example—and the perfectly specifiable process on these objects gets *ipso facto* transmuted into a (Turing) computable function.

Now what is such a coding? It may come cheap to the mathematician, but the cognitive scientist, the psychologist-philosopher sees in it nothing less than a representation. And that—the notion of a representation—is left *completely* out of the picture provided by McCulloch and Pitts on this second reading: there is no trace of "meaning" or "information" left, no account of a mental *state,* nothing but the "engine" driving the mind from state to state.

Of course, this is a typical case of retrospective historiography: the work of the past is seen as an incomplete puzzle, outlining the pieces to be provided by the next generation. Our purpose in this paper being to identify some fundamental differences between paleoconnectionnism and later paradigms, this may perhaps be regarded as an acceptable way to proceed. For us, there is yet another interesting point. Everyone at the time had a clear notion of what was missing: McCulloch and Pitts had only provided an "existence proof" (for any fully specifiable task, there exists a neural net that can accomplish it), and there remained to be built, for all the (important) cognitive tasks,

---

[21] Von Neumann, quoted in Goldstine 1972: 276.

actual nets conforming to the "specifications". The "coding" seemed to pose no problem. Artificial intelligence will take the problem from there, inheriting a costly prejudice. It will in due course discover the paramount importance of coding, in other words, of "knowledge".

2. PERCEPTUAL REPRESENTATION IN PALEOCONNECTIONISM

Actually, it did not escape the first cyberneticians'attention that in one domain of mental activity at least, "existence proofs" and other exhaustivity claims notwithstanding, all the work remained to be done. Not surprisingly, that domain was that of perception, which on one hand had, in the philosophical and psychological traditions, always been regarded as largely independent of cognition (of the higher cognitive processes), and on the other hand raised the problem of representation in a simple form.

What is quite a bit more surprising is to discover that in retrospect, it is only when it approached that problem that connectionism really got underway. The 1943 paper appears to today's connectionists as as false start, and the right track is taken with the much lesser known paper published in 1947 by Pitts and McCulloch (in that order), under the title "How we know universals: The perception of auditory and visual forms", in the same obscure bulletin as their first paper[22]. "It is much more in the direction in which neuroscience and network research has progressed since the 1940's", write the editors of *Neurocomputing* in their introduction. And Rosenblatt, had he known about it (which is unclear as he does not list it among the references to his 1958 paper), would no doubt have approved of it, both because it deals with perception rather than logic, and

[22] Pitts & McCulloch 1947.

because it takes as a major constraint the randomness, unreliability and partial indeterminacy of the nervous system seen at the neuronal level.

The link between the 1943 and 1947 papers is problematic. The second paper makes no mention of the first. But McCulloch, in the grand retrospective he sketches in 1961, clearly claims continuity between the two: "in our next joint paper…"[23], he writes, they showed how to deal with other invariants than time (memory, accounted for with the help of reverberating nets with circles in the 1943 paper, is to be thought of as a "temporal invariant"). Upon examination however, there seems little, if anything, in common between the papers. Neural nets (in the 1943 sense) don't appear, neither do propositions and logic. The single common feature is the appeal to the idea that neurons, or neuronal nets, "compute" values required by the hypothetical account of invariant perception; in other words, the authors operate in the framework of "embodied" computation set up and illustrated in the 1943 paper. As a matter of fact, they attribute to real neurons and neuronal nets properties of formal neurons and neural nets.

These conceptual and/or rhetorical maneuvers should be much more carefully examined than I have the space here and competence now to do. Perhaps we should regard them for the time being as a homage paid by logical vice to perceptual virtue, or by computation to representation. What seems beyond dispute, more seriously, is that the true function of the 1943 paper is to lay the foundations of computationalism, or "embodied" computation, providing both a working hypothesis on the mind/body problem and the technical tools required for the conception of computing devices, "man-made or begotten". This conceptual and technical heritage is claimed by the classical and connectionist paradigms alike. The origins of the schism are to be found elsewhere: in the 1947 paper, for example, and

[23] McCulloch 1961, in McCulloch 1965/1988: 10.

generally wherever the emphasis was placed on perception and neurophysiology. In the actual flow of scientific information and influence, the most salient event relevant to the separation of the two main branches of computationalism came much later: it was the publication in 1958 of Frank Rosenblatt's seminal paper on perceptrons.

The difference in tone between the first McCulloch-Pitts paper and Rosenblatt's is striking. Naturally, the fifteen years separating them, which saw the birth of the cognitive sciences in the contemporary sense and the beginning of their convergence, goes some way towards accounting for the difference. Cybernetics had shown the way, at considerable risk, and the prophet's stance, suited perhaps to McCulloch's situation and personality, would have been quite out of place in Rosenblatt's case. Similarly, clarity and simplicity of expression are easier to achieve in a somewhat settled intellectual landscape. However this cannot be the whole story: the 1947 Pitts-McCulloch paper is closer in tone to Rosenblatt's than to the 1943 paper. At the same time, there is in Rosenblatt's paper an intellectual ambition comparable to the first, rather than the second, of the earlier papers: the focus is on the abstract, completely general characterization of a basic cognitive task, and on the general characterization of devices capable of carrying it out—the 1943 and 1958 papers deliberately situate the main topic on what we would identify today as Marr's top, "computational" level[24]. The 1947 paper is focused on the second, "algorithmic", and third, "implementational", levels.

The basic reason for the disparity, I contend, is that Rosenblatt's topic is easier than McCulloch and Pitts's in 1943 in one crucial respect: representation. Although Rosenblatt has no more of a general theory of representation than his predecessors, he is in a considerably better position than they are because he doesn't need one. To put it crudely,

[24] Marr 1982: 19-31, 336-361.

logic is symbolic, vision isn't[25]. Because duality of levels is essential to logic, and because the relation between levels involves a form of abstract representation, one cannot account for "embodied" logic (or cognitive abilities with a logical dimension) without explicitly postulating an *embodied* representational realm—on pain of actually accounting only for embodied Boolean computation. On the other hand, one *can* set out to account for perception without any prerequisite notion of representation: insofar as the link between a given shape and its "kind" (the class in which it is put by the system under study) can be given a purely (psycho) physical account, so that the "kind" is a *natural sign* of the shape, the job of the theorist is done once he has established the psychophysical (or neurophysiological) chain between a collection of shapes in an environment and their respective "kinds". That job consists precisely in building a theory of a certain kind of concrete representation, and, to repeat, presupposes no general concept of representation.

I thus conclude, with no more than a semblance of paradox, that Rosenblatt's paradigm, just like McCulloch and Pitts's in 1943 (or 1947 for that matter), is computational alright but not representational. In Rosenblatt's case, there simply is no general notion of abstract representation, and this is as it should be. In McCulloch and Pitts's (1943), there should be one and there isn't. But beyond that difference, and the concomitant, obvious one in topics, both theories bear that hallmark of paleoconnectionism: they leave embodied, material representation in the dark.

3. COMPUTATION AND REPRESENTATION IN TODAY'S COMPETING PARADIGMS

[25] That *is* crudely put, for, as is well known, contemporary theories of vision actually favor a largely symbolic conception of vision. As Marr puts it, "the critical point seems to be that even very early vision is a highly symbolic activity" (*op. cit.*: 350). But this, if true, is by no means an *analytic* truth; it is a hard-won empirical conjecture.

In this final section, my goal is to bring out some major differences between paleoconnectionism and neoconnectionism; it will however be useful to stop on the way in order to briefly examine the case of classical computationalism. The pace will be rather brisk, first because apt characterizations of the main contemporary approaches are readily available (see fn. 2), and second because of the retrospective historiography I have, for better or for worse, indulged in: more than once have I characterized a past state of affairs by pointing out what makes it different from our present outlook, or perhaps by describing it in contemporary terms, so that it appeared as a familiar figure with some important feature missing.

a) Classical computationalism

This mainstream approach (sometimes also called cognitivism) in cognitive science and artificial intelligence, from the mid- fifties to this day, rests on a specific way of articulating representationalism, a doctrine about cognitive and mental states, with computationalism, a doctrine about cognitive and mental processes.

Cognitive states in a (von Neumann, AI-programmed) machine are data structures in specified locations; in a human mind, they are inner representations in the functional equivalent of specified locations. In both cases, the medium of the representations or data is an embodied formal language (a "physical symbol system" in Newell and Simon's famous phrase[26]). Formulas of that language are endowed with a semantics, they are information-bearers, for example they express states of affairs or situations. They also have a form and lend themselves to syntactic manipulations.

This is where computationalism comes in. The syntactic manipulations turn out to be effective—they are among the "exhaustively, unambiguously describable «things»" that von Neumann talks about—in other words, they are, up to

[26] Newell & Simon 1976.

coding, computable functions. A Turing machine can therefore effectuate those manipulations, and an *embodied* Turing machine, or any equivalent physical computer, can perform on the *material* aggregates of symbols which constitute the cognitive states of a system the *physical* operations which correspond to the appropriate syntactic transformations. The semantico-syntactic parallelism characteristic of formal languages ensures that appropriate syntactic action leads to appropriate semantic content, so that for example the system is led, in favorable cases, from true beliefs to true beliefs, or from desirable but distant things to still desirable but less distant things.

As already mentioned, McCulloch and Pitts provided, in their 1943 paper, the means to build computer components able to perform elementary computations, and therefore whole computers able to perform any complex (computable) operation—the functional similarity between these components and certain subsystems in the brain could (and can) then be exploited either to model the brain to account for its computational powers, or to build brain-like, or "neuromimetic" computers—but the likeness to the brain is firstly a relative matter, secondly optional. On the other hand, as the other reading of the paper shows, the intention was to apply the computational powers of begotten and man-made devices to propositions. For all these reasons, the 1943 paper heralds classical computationalism at least as much, if not actually more, than neoconnectionism. The absence of a representational medium makes it however a very incomplete prefiguration of classicism, and a close relative of both the Rosenblatt version of paleoconnectionism and also perhaps of certain minority, explicitly nonrepresentational, trends in cognitive science[27].

b) Input/output neoconnectionism

---

[27] See *e.g.* Varela 1979/1989.

Connectionism nowadays is a convenient and somewhat misleading label covering a large variety of doctrines, styles and enterprises, ranging from theoretical biology to psychology, engineering-oriented AI, etc. Of the many important distinctions within this family of research programs, only one need concern us here. A majority of connectionist models are essentially feed-forward neural nets with a distinguished layer of input units and another of output units. Most of the "PDP" (parallel distributed processing) research[28] provides specific models of that kind, as well as a general framework within which to study them, but localist schools[29] also produce input/output information processors. For the present purpose, I therefore propose the nonstandard label "input/output connectionism" as a convenient way of grouping together efforts aimed at producing or describing models of this kind.

These are in many respects the direct heirs to the perceptron[30]. On the other hand, they depart from the perceptron in several important ways.

The best known series of differences bear on the architecture of today's nets and on the almost endless variations affecting the range of activity levels and the transition law of the units, as well as the learning procedure. Nets nowadays have "hidden" units which they use as they please, thus displaying a modicum of self-organization. They can assume a probabilistic character. They follow "back-propagation" or "simulated annealing" learning protocols, etc. These changes (together with enormously more powerful computing resources) have enabled the "neo-perceptrons" to overcome some of the technical

---

[28] Rumelhart & McClelland 1986. The relative numerical importance of PDP makes it tempting to call the whole movement after its main component, but for a number of reasons I wish to resist the temptation.

[29] See *e.g.* Feldman & Ballard 1982.

[30] As pointed out, in particular, by Minsky and Papert in their new fore- and after-words to their classic 1969 treatise on perceptrons.

shortcomings responsible for the virtual disappearance of the original perceptron.

The major conceptual differences lie elsewhere. Input/output connectionism shares with, and in fact for the most part, borrows from, classical computationalism its notion of representation. Admittedly, the emphasis is often laid on the differences: connectionism wishes to reject the classical commitment to some "language of thought" fashioned after the formal languages of logic, it also stresses the advantages and significance of distributed, versus local, representations. But this is not the place to dwell on these differences, and the fact remains that connectionism (of the input/output kind) is representationalist in basically the same sense as classicism[31]. With respect to paleoconnectionism, this is of no little consequence.

By conceiving their nets as embodying processes acting on (embodied) representations, contemporary connectionists extend the range of cognitive phenomena accessible to a connectionistic account or simulation far beyond perception. Just like the classicists, they can help themselves to *abstract,* and not only *natural* representations. The higher cognitive processes can now conceivably be accomplished by nets—while Rosenblatt lucidly admitted that "some system, more advanced in principle than the perceptron, seems to be required at this point."[32] Of course, serious doubts have been voiced about the actual prospects for extending connectionist accounts all the way to language and logical reasoning[33]. Of course, the mere adoption of a general notion of representation is not enough to yield a universal connectionist methodology even in principle applicable to all tasks: in fact, there is considerable disagreement on how the net can be fed the right representation (the critical

---

[31] As stressed, among others, by Fodor & Pylyshyn 1988.

[32] Rosenblatt 1958; A & R: 111, col. 1.

[33] See Fodor & Pylyshyn 1988, Smolensky 1988a, Visetti 1990, Andler 1990, Andler 1990a.

suggestion is that the actual job of assigning the appropriate representations remains outside the scope of connectionism). But these considerations, however deserving of attention in present-day research, do not detract from the importance of the conceptual jump which has allowed neoconnectionism to escape the prison of natural perception: guided by the classical view, it can now approach non-perceptual phenomena in a perception-like way.

As a second consequence, in conjunction with the more specific idea of distributed representations, PDP-style connectionism has developed a two-level, emergentist view of cognition, freeing the physical, computational level from the semantic, symbolic level[34]. This distinguishes PDP connectionism from classicism (at some risk, it must be admitted[35]), and also from paleoconnectionism. Of course, connectionism always distinguished the level of single neurons and local connections from the level of the whole assembly. But paleoconnectionism had no means to link one side of representation to one level, the other side to the other level. This move by PDP connectionism, is most welcome. If successful, it could allow connectionism to finally enter the realm of higher cognitive processes through the front door. Logic at the neuronal level being incompatible with the neoconnectionist view of (formal and neurophysiological) processing, and remaining, as we have seen, but a gleam in McCulloch's and Pitts's common eye, only an emergentist perspective can hope to explain how a net can produce (real) logic. A similar approach is advocated by Smolensky and others for the case of language.

c) Attractor neural networks

Following the physicist John Hopfield, whose 1982 paper was the single most important factor in the renascence of connectionism, and to some extent the psychologist

---

[34] The most thorough defense of this approach is Smolensky 1988.

[35] See Petitot 1990, Andler 1990.

Donald Hebb, whose 1949 book was a major inspiration to paleoconnectionists already, a small group of investigators have begun developing different sorts of nets and using them in an attempt to rethink the most basic concepts of cognitive science or psychology.

No more than a few words can be said about this current, which, being rather less known and in some ways rather more difficult to understand than other brands of connectionism, would require far more space to introduce than I can take here[36]. I will content myself with two hints.

As machines, or physical systems, the nets under consideration (called "attractor neural nets" or ANNs) are distinguished by the fact that their connections are multidirectional (rather than feedforward, as in the input/output schools), and comprise no predefined input or output units. This makes them essentially autonomous dynamical systems, whose behavior can be described (and to some extent predicted) on the basis of their attractor landscapes: in the simplest case, the trajectory of such a system is akin to that of a billiard ball which is initially left to its own device, in a given position and with a given velocity, in a rugged landscape, and finds after a while a line of steepest descent which leads it into a point of local minimal altitude (the simplest kind of attractor), where it rests.

Such a system needs to be interpreted in a novel way in order to be seen as a cognitive device. Hopfield's original idea was to interpret an attractor as a memory (in the psychological—as in "fond memories"… —, not computer science, sense of the word): prompted by some stimulation which places it at a certain location in the landscape, the system eventually stops in an attractor, which is the memory evoked by the stimulus. In a similar vein, Daniel Amit has proposed to interpret the physical event of a rapid return to equilibrium as the correlate of significance: of all the perturbations to which the system is submitted, and launch

[36] The key reference is Amit 1989.

it on a new trajectory, only those which rapidly lead to a new balance are significant, and their meaning or content is the attractor into which they have led the system.

Such a cursory description might suggest arbitrariness or fuzziness in a research program which is singularly devoid of those traits. In fact, the mathematical apparatus brought to bear on the study of ANN's is quite powerful (more powerful by several orders of magnitude than anything McCulloch and Pitts, Rosenblatt or Minsky and Papert could have imagined) and new. Clearly this does not *guarantee* the soundness and depth of the program itself, but it is an auspicious sign. Anyway, its seriousness is unquestionable, and the genuine question that needs to be asked at this point is how do ANNs extend or complete the itinerary which began with McCulloch and Pitts's 1943 paper.

This is my tentative answer. The first cyberneticians had overlooked the notion, and the problem, of representation altogether. Neoconnectionists first focused on the notion and helped themselves, with a hint from the classical computationalists, to a ready-made notion which remained (and remains) to be explicated[37]. Now they seek an account of true, full-blooded, intrinsically intentional representation. Neural nets, concomitantly, started out as Boolean machines, then became non-classical computing devices operating on representations of non-classical medium, and are perhaps in the process of becoming autonomous dynamical systems endowed with intrinsic cognitive abilities dependent on an ecologically valid interpretation.

* * *
*

The very last question may then be whether there remains anything at all of the initial inspiration. Quite a lot

[37] See, *e.g.,* Fodor 1987.

does remain, in my view. First, there is the spirit of the proposed solution to the mind-body problem, based on a mediation by the notion of a machine and that of functional equivalence of different machines—cybernetics'major contribution to the emergence of cognitive science[38]. Second, there is the consideration of "begotten" cognition as the main source of knowledge and constraints: connectionist cognitive science has by and large remained a branch of neuroscience, and has never seriously considered relocating. Third, there is this humble but hardy device, the threshold automaton or formal neuron. Finally, there is the insight that robust collective properties of assemblies of simple components can result from incompletely specified, imperfectly stable, not fully regular or reproducible local characteristics; and that cognition, that robust property of central nervous systems, may thus emerge, phylo and ontogenetically, from a hopelessly complex, ever changing bundle of neurons.

Could this be enough of a reason, as this paper comes to an end, to erase "paleo" and "neo", letting connectionism stretch and progress, through false starts and false endings, from 1943 to this day?

REFERENCES

Amit, Daniel J., 1989. *Modeling Brain Function. The world of attractor neural networks*. Cambridge : Cambridge U.P.

Anderson, James A & Edward Rosenfeld, 1988. **(A & R)** *Neurocomputing. Foundations of Research*. Cambridge, Mass. : MIT Press.

Andler, Daniel, 1986. Studying cognition today, report to the European Science Foundation ; in Eidos **5** : 177-225 ; with two appendices in H. Sinding-Larsen, ed., *Artificial Intelligence and Language*. Oslo : Tano : 201-246.

Andler, Daniel, 1987. Progrès en situation d'incertitude. Le Débat **47** : 213-234.

[38] See Dupuy 1985.

Andler, Daniel, 1990. Connexionnisme et cognition : à la recherche des bonnes questions. *Revue de Synthèse*, série générale tome CXI, **IV** #1-2 : 95-127.

Boolos, George S. & Richard C. Jeffrey, 1974. *Computability and Logic*. Cambridge : Cambridge U.P.

Clark, Andy, 1989. *Microcognition. Philosophy, Cognitive Science, and Parallel Distributed Processing*. Cambridge, Mass. : MIT Press.

Cummins, Robert, 1989. *Meaning and Mental Representation*. Cambridge, Mass. : MIT Press.

Dupuy, Jean-Pierre, 1985. L'essor de la première cybernétique (1943-1953). *Cahiers du CREA* **7**. Paris : Ecole polytechnique.

Feldman, Jerome A. & Dana H. Ballard. Connectionist models and their properties. *Cognitive Science* **6** : 205-254. Repr. in A & R.

Fodor, Jerry A., 1975. *The Language of Thought*. New York : Thos. Crowell, 1975 ; repr. Cambridge, Mass : Harvard U.P.

Fodor, Jerry A., 1981. *Representations : Philosophical Essays on the Foundations of Cognitive Science*. Cambridge, Mass. : MIT Press. A Bradford Book.

Fodor, Jerry A., 1983. *The Modularity of Mind*. Cambridge, Mass. : MIT Press. A Bradford Book.

Fodor, Jerry A., 1985 Fodor's guide to mental representation : The intelligent Auntie's vade-mecum. *Mind*. **XCIV** : 76-100.

Fodor, Jerry A., 1986. Information and association. *Notre-Dame Journal of Formal Logic*, **27** : 307-323.

Fodor, Jerry A., 1987. *Psychosemantics. The Problem of Meaning in the Philosophy of Mind*. Cambridge, Mass. : MIT Press.

Fodor, Jerry A & Zenon Pylyshyn, 1988. Connectionism and cognitive architecture : A critical analysis. *Cognition* **28** : 3-71.

Gardner, Howard, 1985/1987. *The Mind's New Science : A History of the Cognitive Revolution*. New York : Basic Books.

Goldstine, Herman H., 1972. *The Computer from Pascal to von Neumann*. Princeton : Princeton U.P.

Graubard, Stephen R., ed. 1988. *The Artificial Intelligence Debate. False Starts, Real Foundations*. Cambridge, Mass. : MIT Press. A special issue of *Daedalus* **117** #1 (Winter 1988).

Grossberg, Stephen, 1980. How does the brain build a cognitive code ? *Psychological Review* **87** : 1-51. Repr. in A & R.

Haugeland, John 1981, ed. *Mind Design*. Cambridge, Mass. : MIT Press.

Haugeland, John, 1978. The nature and plausibility of cognitivism. *Behavioral and Brain Sciences* **1** : 215-226 ; repr. in Haugeland 1981.

Hebb, D.O., 1949. *The Organization of Behavior*. New York : Wiley.

Hebb, D.O., 1980. *Essay on Mind*. Hillsdale, N.J. : Lawrence Erlbaum Associates.

Hinton, Geoffrey E. & James A. Anderson, 1981. *Parallel Models of Associative Memory*. Hillsdale, NJ : Erlbaum.

Hopfield, John, 1982. Neural networks and physical systems with emergent selective computational abilities. *Proc. Natl Acad. Sc*. USA **79** : 2554-2558. Repr. in A & R.

Lévy, Pierre, 1985. Analyse du contenu des travaux du Biological Computer Laboratory. *Cahiers du CREA* **8**. Paris : Ecole polytechnique : 155-191.

Livet, Pierre, 1985. Cybernétique, auto-organisation et néo-connexionnisme. *Cahiers du CREA* **8**. Paris : Ecole polytechnique : 105-153.

Livet, Pierre, 1990. Second cybernetics : A double strategy for representing cognition, *Communication and Cognition*, **23**, #2/3, special issue on « The old and new in cybernetic fashions », G. Van de Vijver, ed. : 213-221.

McClelland, James L., David E. Rumelhart, & the PDP Research Group, 1986. *Parallel Distributed Processing : Explorations in the Microstructure of Cognition*. Vol. 2 :

Psychological and Biological Models. Cambridge, Mass. : MIT Press. A Bradford Book.
McCulloch, Warren S., 1965/1988. *Embodiments of Mind*. Cambridge, Mass. : MIT Press.
McCulloch, Warren, 1961. What is a number, that a man may know it, and a man, that he may know a number ?, *General Semantics Bulletin*, # 26-27 : 7-18. Repr. in McCulloch 1965/1988 : 1-18.
Minsky, Marvin & Seymour Papert, 1969/1989. *Perceptrons*. Cambridge, Mass. : MIT Press.
Newell, Allan, 1983. Intellectual issues in the history of artificial intelligence, in F. Machlup & U. Mansfield, eds, *The Study of Information : Interdisciplinary Messages*. New York : Wiley.
Newell, Allen & Herbert A. Simon, 1976. Computer science as empirical enquiry : Symbols and search. *Comm. Am. Ass. Computing Machinery* **19** : 113-126. Repr. in Haugeland 1981.
Petitot, Jean, 1990. Le Physique, le Morphologique, le Symbolique : remarques sur la vision. *Revue de Synthèse*, série générale tome CXI, **IV** #1-2 : 139-183.
Pinker, Steven & Jacques Mehler, 1988. *Connections and Symbols*. Cambridge, Mass. : MIT Press. A Bradford Book.
Pitts, Walter & Warren S. McCulloch, 1947. How we know universals : The perception of auditory and visual forms. *Bulletin of Mathematical Biophysics* **9** : 127-147. Repr. in McCulloch 1965/1988 and in A & R.
Pylyshyn, Zeno, 1984. *Computation and Cognition. Toward a Foundation for Cognitive Science*. Cambridge, Mass. : MIT Press. A Bradford Book.
Rosenblatt, Frank, 1958. The perceptron : A probabilistic model for information storage and organization in the brain. *Psychological Review* **65** : 386-408. Repr. in A & R.
Rumelhart, David E., James L. McClelland, & the PDP Research Group, 1986. *Parallel Distributed Processing : Explorations in the Microstructure of Cognition*. Vol. 1 :

Foundations. Cambridge, Mass. : MIT Press. A Bradford Book.

Shannon, Claude E., 1938. A symbolic analysis of relay and switching circuits, *Trans. AIEE*, **57** : 713 sq.

Smolensky, Paul, 1988. On the proper treatment of connectionism. *Behavioral and Brain Sciences* **11**: 1-37.

Smolensky, Paul, 1988a. Connectionism, constituency, and the language of thought, in B. Lœwer & G. Rey, eds. Jerry Fodor and His Critics (forthcoming).

Varela, Francisco J., 1979/89, *Principles of biological Autonomy*, Elsevier/North Holland, New York ; version fr. *Autonomie et connaissance*, Paris : Le Seuil.

Visetti, Yves-Marie, 1990, Modèles connexionnistes et représentations structurées, *Intellectica* **9-10** : 167-212.