INTRODUCTION TO VOLUME I

Jerome Y. Lettvin

At the start of the 20th century in the United States, psychology reflected none of the epistemological problems outlined in the 19th century, none of the metaphysical questions raised by the existence of these problems, and certainly no clue as to how to go from structure to function in the case of the nervous system.

Let me illustrate the point. If one were to get a book on the kidney that never once mentioned urine or the production of urine, one would feel cheated. Or if one had a long essay on muscle which never mentioned contraction, one would certainly turn the essay back. Yet one could publish a long treatise on the anatomy and physiology of nerve, on the diagnostic criteria for nervous disease, and never once consider mental process. Against this background, not particularly rich, the background of William James, of the English philosophers - Russell included, there was no possible approach to a theory of nervous action, since a theory would imply: things are this way for such and such a reason. But the notion of reason as we now see, was read out of natural science by the English philosophers.

To this very day that rather dreadful state of affairs has prevented neurosciences from exhibiting any theory whatsoever. As the data in the field grow to the point where no man can read it all, much less comprehend it, there is no background, no backbone, no structure or theory against which to understand the details. To my knowledge, the McCulloch-Pitts theory of the brain is the only one that has ever been issued, and it, itself, was issued more or less to the indifference of those who most needed it.

It does not matter so much whether theory is right or wrong, just that there be a theory, not simply a vague hypothesis. The McCulloch-Pitts theory of nervous action remains to date the only one available for nervous action. It is certainly wrong, but in its wrongness, just as Bohr's wrongness in his view of the atom, it contains the seeds of a new theory.

At present, the McCulloch-Pitts theory, in one way or another, becomes the foundation for the field called Artificial Intelligence, which itself is an embattled attempt to provide a theory relating brain and mind.

Let me dwell a little on the philosophical posture that was held by the sciences of the nervous system between 1900 and 1943. To do this, consider the history of that posture. The rules of natural science laid down in the 17th century from the time of Galileo to the time of Leibnitz and Newton were that the external world is given in terms of magnitudes, figures (or, if you wish, arrangements) and motions or changes. So that what was available to the observer were not the objects of his perception but things that could be interpreted by the observer to be objects and their negotiations through a world. The function of science was to relate the measurements made on observables to each other such that given a set of measures one could write out the laws by which observables changed. So far as the observer goes, he has perceptions, knowledge, memories, things of this sort, which are not measurable under any circumstances because they are private and inaccessible. To relate measured observables to those things which are not measurable is impossible.

In a sense, a theory of mind could not be a theory in natural science. Yet, from Aristotle on, and very much regarded by Newton as well as by the Continental philosophers, was the tradition that nothing could be had in the knowledge of the observer in terms of content, empirical content, that was not initially in the senses. This, of course, is the point which is made by Leibnitz, and then by Kant. Were the observer not so bound to observables science itself would be impossible, since an observation would have no strict basis.

What relates that which is observed to the processes in the observer must be something akin to information rather than to the energetics of the world and, although Leibnitz had made an attempt at it, information had not yet been clearly defined. If one were given axioms by which sense data, which still are physical, could somehow or other be related to the state of mind in the observer, then and only then would a theory of observation or a theory of knowledge be possible, having at least one foot in the natural sciences.

It is one thing to say that all the source of empirical knowledge lies in the senses, and another to take the consequences of that. Curiously enough an American engineer/scientist, J. Willard Gibbs, uttered what is to be regarded as the fundamental epistemological law by which it might be possible to go from physiology to mind.

Gibbs was one of McCulloch's heroes. He was one of the few native American philosopher/scientists; unfortunately not sufficiently well-regarded by the Americans themselves. Gibbs' phase rule*, issued at a time when it was supposed to apply only to thermodynamics, had an epistemological significance which was noted at that time by Gibbs' European colleagues, but was ignored back home. The principle laid down by Gibbs in his famous rule describes what one can say by virtue of the information given, and it lays an important constraint upon what can be perceived given the sense data. It does not tell you specifically what can be perceived but what it is impossible to perceive given those data. An additional principle, corollary to the phase rule, and only expressed clearly by the time of Shannon (and implicit in a good deal of work up to the time of Shannon) is that information which is lost in process is forever lost and cannot be supplied to later parts of a process by such things as revelation. Nobody has ever bothered to make such rules as Gibbs' rule and the information loss rule -- axiomatic in such a way as to make it possible to

* Gibbs' phase rule brings into natural science the formal notion that if a system is to be defined in terms of a state, every one of the ways in which that system can vary from that state must be constrained. In short, if I have a system defined by N variables, there must be n-independent equations needed to provide a specific solution to those equations for each of the variables.

proceed, if not directly to perception, at least to the statements of what perception cannot be.

In the heritage of McCulloch and Pitts, there was an additional factor. Leibnitz in the 17th century had designed, although he had not been able to build, the first logical machine. This was not known until the 1950's, when interest in logical machines began again. Leibnitz had said of such a machine that in the future when philosophers disagree they will not fight with each other but say to each other, "Let us sit down and compute."

The history of computing machines is reasonably well known from the time of Babbage on, that is, for a century and a half after Leibnitz. The possibility of logical machines, namely, devices that would be able to compute any computable number, was very much in the air as early as the 1940's. At that time Turing's paper on general treatment of all logical machines was available more or less as a curiosity in mathematical circles and relatively unknown in the biological community. Quite independently, McCulloch and Pitts set about looking at the nervous system itself as a logical machine in the sense that if, indeed, one could take the firings of a nerve fiber as digital encoding of information, then the operation of nerve fibers on each other could be looked at in an arithmetical sense as a computer for combining and transforming sensory information.

There was a good deal of reason for their notion. First, nerve impulses were all-or-none in character; that is, at any instant there would be a pulse or no pulse in a particular place along a nerve fiber. Second, the rules for combination had already been explored in part so that it was known that one could prevent firing of cells by afferent pulses from other cells and one could also promote firing by other afferent pulses.

The concepts of inhibition and excitation in the nervous system go back to the 19th century and reach their English fruition in The Integrative Action of the Nervous System by Sherrington, written in 1911. In the 1940's, inhibition and excitation on a monosynaptic basis had already been established by the remarkable work of David Lloyd. This profoundly impressed

McCulloch. It was, in fact, the examination of the consequences of inhibition and excitation upon single motorneurons that led McCulloch and Pitts to the supposition that one could apply such principles to advantage elsewhere in the nervous system. And so they set about laying out the structure of the nervous system as if it were a network of gates, exactly as in current computers, in order to compute particular functions from sense data.

Implicit in the McCulloch-Pitts design were the two notions mentioned earlier; first, that all of the data on which the content of knowledge is founded is provided by the senses and is given form by the structure of the system acting as the embodiment of the synthetic *a priori*, the processor; second, that any information lost in the process is permanently lost. This is given in the nature of the design of the neurons such that they can only operate on information received in the same layer.

Because this theory is, as I say, the only extant theory, however wrong it may be, of the relation of brain to mind, it is very useful to trace the precursors of that theory as it developed in the mind of McCulloch. His personal history had been given earlier, but I want to retell parts of it in regard to the points that I want to make.

Early in his life, McCulloch became interested in the metaphysics of Emmanuel Kant, and in particular was very much taken with the problem of understanding the notion of the synthetic *a priori*. That kind of knowledge, which itself was not informative, gave form to the data or, if you wish, processed the sense impressions which are the connections with the external world. The notion that the synthetic *a priori* was a kind of processor was already implicit in the way Kant expressed it. The notion of giving form to that program became an obsession with McCulloch. It is one thing to utter a kind of general principle underlying a process in terms of what a mechanism can provide. For Warren, in a much more explicit way than for almost any other of his contemporaries, the brain was strictly a mechanism. It was not something that would ultimately remain mysterious, but would have to operate by virtue of rules. And he was going to find those rules somehow or other. But no matter how one regarded the problem of the brain/mind relation, perceptions, knowledge, memory, all of these mental functions remained so undefined that it was impossible to say what kind of structure would lead to them in a believable way. It occurred to Warren and Walter working together that one of the ways by which data can be manipulated and given form is to encode them, which subjects them to logical operations in a machine.

Walter Pitts, who was companion, protégé and friend to Warren, had, for a long time, been convinced that the only way of understanding nature was by logic and logic alone. Up to the early 1940's, McCulloch's thinking was rather vague, as it had to be before the actual issues took shape. He knew that he wanted a kind of nervous operation that would do useful things, but the fundamental question remained -- what was it that a neuron could do? Although he knew his logic thoroughly, he did not regard it in the same impassioned way as did Pitts. And although he knew of the work of Boole and knew very well that the Boolean logic could be applied to some mental processes, the notion of embedding that logic in a neuronal structure occurred only after the collaboration began. That was because Pitts had committed himself to logic as the key to the structure of the world in a way that no other person that I know had ever done.

When you try to think of alternative manners of handling data, except by program, you are hard put to imagine them. The general rules laid down by Kant about the synthetic *a priori* are not specific about either data from the outside world, the empirical synthetic aspect, or about the program that handles those data. Since logic is the only successful method we know to handle data in general, and since all natural historical theory can, in one way or another, be reduced to logical manipulation, it was inevitable that the representations that McCulloch and Pitts both wanted and were successful in obtaining were representations in terms of a logical machine. In fact, what they had done in 1943 was to achieve ahead of time a kind of program for handling data before computers even existed. Strongly in the minds of both McCulloch and Pitts were the notions of Russell as contained in his essays on mind, the notions of Peirce, and to a great extent, the notions of Whitehead, in particular as regards the structure of mind and experience. It was inevitable, therefore, that they should deal with the brain as a logical machine. But certain personal experiences of McCulloch made this an even stronger image than he would have professed on logical grounds alone. His experimental researches with Dusser de Barenne on the strychnine localizations in the brain, and his own vast reading of the work done in the control of motor system, led him to the notion of a structure that must, in its internal working, be logical.

Let us regard, for example, the work done following the lines laid down by Dusser de Barenne. On strychninizing a single patch of cortex, he was able to show that the lines of communication, that is, the axonal structures leaving that portion of cortex, led to very specific other cortical points. And the way he was able to tell this was by the production of a massive, single synchronous volley occurring in the place strychninized and proceeding as a recordable volley to other points in the way that pulses would go down telegraph lines. Since the only thing that could travel down those lines were the pulses that originated in the strychninized region, then these pulses carry the information that would issue from that region to other regions. And so he had a vast account of what happened behaviorally when different portions of the brain were strychninized so that impulses proceeded from there to elsewhere. And he would give vivid accounts of cats with strychnine applied to a particular part of the brain turning and biting and scratching at particular regions on their bodies there represented, indicating that something of a sensation was set up by these synchronous pulses. Similarly he would observe, in common with his other physiological colleagues, that if one stimulated a particular nerve or a particular portion of the brain, the stimulus there, although electrical in nature, not coming from the external world but set up in the substance of the brain itself, would be attended by definite and very vivid kinds of experience. This kind of thinking was profoundly reinforced by Penfield's observations of particular, definite percepts attained from electrical stimulation of one area of cortex. Similar were the observations of Percival Bailey, who was one of McCulloch's very good friends. So too were the observations made by a variety of students on specific auras connected with epileptic seizures. So that, although on no account would Warren subscribe to a jukebox theory of the brain (for which he parodied Walsh, the British neurologist), nevertheless there was no question in his mind that the pulses that moved down pre-existing paths from one place to another, acting inhibitorily or excitatorily, were responsible for all the kinds of perception, thinking and memory that we enjoy. If these pulses could be expressed as all-or-none entities, then one might consider that the language with which the brain talks to itself consists of strings of zeroes and ones exactly as one would have in a digital computing device built on a binary system. So, it was inevitable that one should take the easiest and most perspicuous way of devising a computer as a model of the brain.

A second feature of his work with Dusser de Barenne also played a strong part. That was the notion of the irreversibility of the synapse. That is to say, information could proceed in only one direction given a system in which synapses were the copulae. So that while it is possible that one can always build a cell that is self-excitatory or self-inhibitory by having axonal branches ending back on itself, these exotic kinds of elements, useful in a formal way, only make more poignant the specific idea that information goes in only one direction through a nervous net.

It would be impossible to devise a logical system in which the connections were reversible; that is, active informationally in both directions. So, to McCulloch's mind, the existence of a single direction in the nervous system for information reinforced the idea of an essentially logical device.

Yet, early in their thinking, both he and Pitts, looking at the material, conceived the notion that two-valued logic is not sufficient. In a word, while their machines worked on a relatively low level, the contingent aspect of experience and perception was not something that they could afford to ignore. Shortly after the paper on the logical calculus, McCulloch produced another dealing with the heterarchy of values in a nervous system. The hierarchical structure implied by the machine that he and Pitts had devised did not seem adequate to experience. In that paper, *The Heterarchy of Values*, he explored the contingent aspect of perceptions by dealing with circularities of preference. This one paper by itself set him to thinking of systems of logic in which one did not have a simple yes-no, one-zero kind of element -- the common gate -- but, instead, one in which whether an element occupied one state or the other was contingent not only on the immediate information coming to it but on stored information as well. He sought some way to bring memory into play in a way that would not be so exaggerated as to be unrealizable, which would be the case if he pursued memory by a kind of logical net structure. The notion of many-valued logic was already given by a variety of writers in the field, although it never entered biology, but it was a proposed way of handling some formal propositions in the field of logic itself.

Yet a third influence played upon the development of the logical calculus, and that was the extreme regularity, however complex it appeared in the sketching, that was noticed by Ramón y Cajal in the descriptions of the neurons of a particular region. While to the uninitiated it seems as if the nervous system, as shown by Golgi stain, is complex beyond any reason, there is a certain repetitive order in structure that is not trivial. Indeed, the way Ramón y Cajal drew his illustrations makes the point. Ramón y Cajal never drew cells directly by looking at them through a microscope. What he did was to look at a particular region of the brain for several weeks, if necessary, day after day, element after element, and then one day he would close up his instruments, sit down and draw what he had seen, thus abstracting what might be called the visible invariants of the tissue without giving any specific cell or specific connection a hegemony, unless it were so often repeated that he could not miss it. These invariant structures inside a particular portion of the brain, and for that portion of the brain invariant over all of the animals of that species, led McCulloch and Pitts to the notion that the structure of the brain dictated the logic. And it is a point that to this day is not easily controverted. It is a synthetic, indeed an esthetic task to take such an assemblage of neurons, complexly interconnected but of repetitive structure, and read into the design a function; a very dangerous step but a most useful one and certainly better than a professed ignorance.

The attitude of McCulloch and Pitts to the complexity, at which others turned up their hands and backed away, is perhaps best given by an anecdote which does not concern either of them. In the 19th century there was a neurophysiologist by the name of Dubois Raymond who would travel the lecture circuits of Berlin (for this was before television) with a nicely prepared lecture about the nervous system and its functions. He ended with the ringing motto: *Ignoramus et ignorabimus*, "we don't know and we won't know." So incensed was the mathematician Hilbert by this horrible motto that he caused inscribed on his tombstone, *Wir müssen wissen, und wir werden wissen*, "we must know and we will know." And it is delightful to note that McCulloch and Pitts in this respect were on Hilbert's side.

The influences that played on McCulloch and Pitts for the construction of their paper have been recounted. But what confirmed them in their belief that the notions they had were not irrelevant was the move by von Neumann into the construction of a logical machine electronically run realizing the dreams of Babbage. Once such a machine was possible, and this occurred in only a few years after the appearance of their paper, the great temptation was to suppose that now, by means of such a machine, one finally would be able to model the brain. This certainly was McCulloch's hope. But Pitts, at this point, began to dissent. In 1949 von Neumann brought out an essay, *The Natural and Logical Theory of Automata*, at the Hixon Symposium, in which, after paying due tribute to the McCulloch-Pitts *Logical Calculus*, and recognizing it as a theory, he nevertheless took it to task as insufficient to account for experience. According to von Neumann's criticism, the categories of experience cannot be such as those to which known logic applies. The paper is interesting because it not only deals with the McCulloch-Pitts theory of the brain, but also deals with the notion of self-replication of such systems and announces, three years prior to the discovery of the DNA basis for genetics, what the structure of genetics must be. McCulloch and Pitts took the paper seriously but could not respond to the criticism any more than von Neumann himself could show a way out. Von Neumann was of the opinion that the logical design of his computer would indeed make all sorts of computation possible, but only under the specification that you knew exactly what you wanted; that is, it could compute any computable number in exactly the same way as could the

McCulloch-Pitts nerve net. The question that von Neumann raised not only about his machine but about the McCulloch-Pitts model was whether or not this was the essence of perception, thinking, memory and the like; that you were computing specific computable numbers or something that could be mapped onto them.

Later on, Minsky, who took his doctorate under von Neumann, was to say that this was an aberration of von Neumann's. That is to say, it was a confession of weakness on von Neumann's part, because he had not enough faith in the structure that he had built. But this, I think, is a way of avoiding the issue raised by von Neumann as much against himself as against the McCulloch-Pitts theory.

One would assume, I think, that the presence of a theory, however strange, in a field in which no theory had previously existed, would have been a spur to the imagination of neurobiologists, if I may use so horrid a term. But this did not occur at all! The whole field of neurology and neurobiology ignored the structure, the message, and the form of McCulloch's and Pitt's theory. Instead, those who were inspired by it were those who were destined to become the aficionados of a new venture, now called Artificial Intelligence, which proposed to realize in a programmatic way the ideas generated by the theory. In no sense was Artificial Intelligence going to resemble or explain natural intelligence. Instead, it would show that monsters that could, in fact, act like humans or like animals could be built and, therefore, could be used as representations of the original. This field, Artificial Intelligence, so peculiarly engendered by McCulloch, was one, for some reason, McCulloch himself avoided. In retrospect, it is hard to say why. Minsky's work was not inconsiderable. Minsky and Papert between them had already begun designs on game-playing machines -- task-oriented machines that were to become in the short period of a decade or two more impressive. Not that they ever realized anything so simple as a perception, but they did tasks that were formerly thought to be peculiarly human. For example, they played chess, they piled blocks on each other, they did what two or three year old children are said to have to learn to do, and in a way represented exactly the sort of caricature necessary before one begins a filling-in job. Yet, with all of this promise held forth by AI, McCulloch could not himself take it as seriously as he might have, and that was because he felt that a logical machine by itself, however cleverly programmed, would come to naught. It was very hard to find out from him why this was the case. He staunchly supported Minsky and Papert in their endeavors, and indeed they readily agree that they took inspiration from McCulloch. Yet something lay in McCulloch's mind, saying this is not the case; that is, there must be another kind of model that should be more suitable. And he began playing with all manner of logic. I mentioned the three-valued logic which arose from the paper *A Heterarchy of Values*. He also played with probabilistic logics. He played with all manner of strange notations to try to find one that would appear to him consonant with what it was he felt neurons could do. It is hard not to sympathize with him, because it is better to be Don Quixote than Sancho Panza in any view of the world. That is, it is better to go with a lance after a dream than to accept the status quo, however profound the status quo may seem.

And so, for a while he went back to the philosophers who initially had informed him, and again he restudied Leibnitz, Kant, Peirce, the various schools which they set up and others set up after them, and I well remember his revulsion when he came to Hegel reconsidered. There was no nonsense to his tongue-in-cheek view of their considerations of epistemology. He never seemed to be able to find a philosopher to whom he would accord an insight into the nature of the problem that he was attacking. One line of thought he found immensely attractive. In his reaction against Freud, seeing in Freud the simple repetition of what was in Plato's Republic, the notion occurred to him of a command structure in the nervous system, almost military, certainly governmental, possibly naval in structure; a system of command and control by which experience as much as intention and all those other aspects of epistemology could be formed. He had investigated the reticular formation in the brain stem, together with his colleagues Magoun and Snider at Northwestern University. This region anatomically described by Paul Yakovlev as "the original beast", the fundamental animal structure of the brain, seemed to Warren to be not necessarily the place of the ego, but certainly the place whence all of the rest of the brain was organized in some instructional way. He was delighted when Percival Bailey, touching the opening of the aqueduct from the third ventricle, showed that a patient instantly went to sleep, and called that region "the center of unconsciousness", which is a delightful parody. In almost all the instances of encephalitis, wherein the circum-aqueductal structure was involved, there was profound disorder in thought, in consciousness, and in ability to perform in spite of the fact that the number of cells in this lining of the aqueduct and ventricles is relatively small compared to the number of cells in the rest of the nervous system. It was important that Magoun was able to show that he could either activate an animal -- that is, arouse an animal from sleep -- or put an animal to sleep by stimulating in different portions of the reticular formation. In those parts of the reticular formation called the magno-cellular groups, there lies the capacity of either inhibiting or exciting the whole sphere of action via long tracts to the spinal cord. I remember how excited he was when Oliver Selfridge and I one night showed that, in an animal intoxicated with a fatal dose of strychnine, we could prevent all strychnine convulsions by continuous stimulation of the bulbo-reticular inhibitory tract. It seemed to him, as it seemed to Paul Yakovlev, and as it seemed to Magoun and his cohorts, that the control of the whole nervous system lay in this innominately connected reticulum around the central canal of the brain and spinal cord. He switched, therefore, from examining the input and how it was to be processed to examining the control system that ordered the processing. While it would be difficult to say exactly how the reticular formation as a biological structure would engage the whole brain, throwing it into one state or another, that it do so seemed to be unarguable to McCulloch, and certainly in many respects seemed to follow from the physiological experiments themselves.

Accordingly he transferred his attention from the relatively intractable problems of dealing with perception and memory to dealing with control processes, or the control of processes that themselves were the synthetic *a priori* needed by Kant. In this way, which represented a kind of impossible dream by his own account, he was joined by a variety of other researchers. Here the data were not to be had, and it was not that states of the reticular formation could be represented in such a way as to be experimentally accessible. It was a set of gedanken experiments entirely devised to give a kind of eminence grise that chose between synthetic a prioristic programs. This is the first time that McCulloch departed from a base in empirics. He was going to try to devise the system completely divorced from the possibility of any measurement, and devise it in such a way that the logic used could be applied to any control system. It was an ambitious program and one that is very hard to understand.

The notion of a control system somewhat independent of what is to be controlled is like a Cheshire cat's smile. Yet, in a way, the experiments of Magoun and Snider in which Warren also collaborated, could not help but give the impression that such a system must exist. And accordingly, Warren set about how to envisage an optimum naval command as a way of looking at the fundamental governing structures of nervous activity. Not wishing to give the notion that he was designing for military purposes, I mean only that he was looking at general principles involved in a hierarchical structure of command.

Even in this, however, he was to be frustrated, because the handling of such a system as he wanted required minimax operations of a kind which had not yet been realized. One of the difficulties in dealing with representations of systems is that singularities which are, in fact, an important part of a theory of control in the real world are very difficult to represent formally. The notion of a general mode of handling such singularities, therefore, was quixotic in the extreme. And yet, as Walter Pitts once remarked, there are only two kinds of problems: trivial and insoluble, and an insoluble problem becomes trivial once you have solved it. It was certainly McCulloch's character that the more insoluble the problem seemed the more he was likely to attack it.