GUALTIERO PICCININI

THE FIRST COMPUTATIONAL THEORY OF MIND AND BRAIN: A CLOSE LOOK AT MCCULLOCH AND PITTS'S "LOGICAL CALCULUS OF IDEAS IMMANENT IN NERVOUS ACTIVITY"

ABSTRACT. Despite its significance in neuroscience and computation, McCulloch and Pitts's celebrated 1943 paper has received little historical and philosophical attention. In 1943 there already existed a lively community of biophysicists doing mathematical work on neural networks. What was novel in McCulloch and Pitts's paper was their use of logic and computation to understand neural, and thus mental, activity. McCulloch and Pitts's contributions included (i) a formalism whose refinement and generalization led to the notion of finite automata (an important formalism in computability theory), (ii) a technique that inspired the notion of logic design (a fundamental part of modern computer design), (iii) the first use of computation to address the mind–body problem, and (iv) the first modern computational theory of mind and brain.

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... 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 Pitts'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 (Lettvin 1989a, 17).

Warren S. McCulloch and Walter H. Pitt's 1943 paper, "A Logical Calculus of the Ideas Immanent in Nervous Activity," is often cited as the starting point in neural network research. As a matter of fact, in 1943 there already existed a lively community of biophysicists doing mathematical work on neural networks.[1] What was novel in McCulloch and Pitts's paper was a theory that employed logic and the mathematical notion of computation – introduced by Alan Turing (1936–37) in terms of what came to be known as Turing Machines – to explain how neural mechanisms might realize mental functions. The present paper reconstructs McCulloch and Pitts's intellectual context, elucidates their theory of mind and brain, and argues that their contributions included (i) a formalism whose

Synthese **141:** 175–215, 2004.

refinement and generalization led to the notion of "finite automata" (an important formalism in computability theory), (ii) a technique that inspired the notion of logic design (a fundamental part of modern computer design), (iii) the first use of computation to address the mind–body problem, and (iv) the first modern computational theory of mind and brain.

McCulloch and Pitts's theory is *modern* computational in the sense that it employs Turing's mathematical notion of computation. So for instance, although Kenneth Craik's theory of mind and brain was published at roughly the same time (Craik 1943), it is not a *modern* computational theory in the present sense because it appeals to computation only in an informal sense. The modern computational theory of mind and brain is often credited to Turing himself (e.g., by Fodor 1998). Indeed, Turing talked about the brain first as a "digital computing machine,"[2] and later as a sort of analog computer.[3] But Turing made these statements in passing, without attempting to justify them, and he never developed a computational theory of thinking. More importantly, Turing made these statements well after the publication of McCulloch and Pitts's theory, which Turing knew about.[4] Before McCulloch and Pitts, neither Turing nor anyone else had used the mathematical notion of computation as an ingredient in a theory of mind and brain. The present paper aims, among other things, to point out how McCulloch and Pitts's theory changed the intellectual landscape, so that many could see neural computations as the most promising way to explain mental activities.

I will argue that McCulloch and Pitts's computational theory rested on two principal moves. First, they simplified and idealized the known properties of networks of neurons so that certain propositional inferences could be mapped onto neural events and *vice versa*. Second, they assumed that individual neural pulses had propositional contents that directly explained mental processes. Neither of these moves is likely to find supporters today, at least not in the form proposed by McCulloch and Pitts. And yet many contemporary authors profess to agree with McCulloch and Pitts that brains perform computations, and that neural computations explain mental activities. Contemporary computationalists may be interested in studying how computationalism was initially justified, and in pondering whether their version of computationalism is better justified than McCulloch and Pitts's version in terms of the known properties of neural mechanisms.

In spite of its importance, McCulloch and Pitts's paper is often misrepresented. For instance, a common misconception is that McCulloch and Pitts demonstrated that neural nets could compute anything that Turing Machines could:

> McCulloch and Pitts proved that a sufficiently large number of these simple logical devices, wired together in an appropriate manner, are capable of universal computation. That is, a network of such 'lineal threshold' units with the appropriate synaptic weights can perform any computation that a digital computer can, though not as rapidly or as conveniently.[5]

As we shall see, this is incorrect in two respects. First, McCulloch and Pitts did not *prove* any results about what their nets could compute, although they claimed that there were results to prove. Second, McCulloch–Pitts nets – as McCulloch and Pitts explicitly recognized – were computationally less powerful than Turing Machines.

McCulloch and Pitts's theory raise many conceptual issues. The great historical and philosophical significance of their paper, as well as its common misrepresentation, warrant that we devote some close attention to it.[6]

1. TOWARDS A MECHANISTIC THEORY OF MIND

Although McCulloch had a keen interest in philosophy and mathematics, in which he took several undergraduate and graduate courses, he was mainly a neurophysiologist and psychiatrist. He believed that the goal of neurophysiology and psychiatry was to explain the mind in terms of neural mechanisms, and that scientists had not seriously tried to construct a theory to this effect.[7]

While pursuing his medical studies in the mid-1920s, McCulloch claimed that he developed a psychological theory of mental atoms. He postulated atomic mental events, which he called "psychons," in analogy with atoms and genes:

> My object, as a psychologist, was to invent a kind of least psychic event, or "psychon," that would have the following properties: First, it was to be so simple an event that it either happened or else it did not happen. Second, it was to happen only if its bound cause had happened... that is, it was to imply its temporal antecedent. Third, it was to propose this to subsequent psychons. Fourth, these were to be compounded to produce the equivalents of more complicated propositions concerning their antecedents.[8]

McCulloch said he tried to develop a propositional calculus of psychons. Unfortunately, the only known records of this work are a

few passages in later autobiographical essays by McCulloch himself.[9] The absence of primary sources makes it difficult to understand the nature of McCulloch's early project. A key point was that a psychon is "equivalent" to a proposition about its temporal antecedent. In more recent terminology, McCulloch seemed to think that a psychon has a propositional content, which contains information about that psychon's cause. A second key point was that a psychon "proposes" something to a subsequent psychon. This seems to mean that the content of psychons can be transmitted from psychon to psychon, generating "the equivalents" of more complex propositions. These themes would play an important role in McCulloch's mature theory of the brain.

McCulloch did his internship in organic neurology under Foster Kennedy at Bellevue Hospital in New York, where he finished in 1928.[10] While working as an intern, he "was forever studying anything that might lead me to a theory of nervous function."[11] He developed a long-term interest in closed loops of activity in the nervous system, that is, activity flowing through neurons arranged in closed circuits. Since neural activity flowing in circles along closed circuits can feed itself back onto the circuit, thereby sustaining itself indefinitely, McCulloch called this process "reverberation." At that time, there was no evidence of closed anatomical loops within the central nervous system, although McCulloch attributed to Ramón y Cayal the hypothesis that they exist.

The tremors of Parkinson's disease, McCulloch thought, could be explained by closed loops of activity connecting the spinal cord and the contracting muscles. With his fellow intern Samuel Wortis, McCulloch discussed whether the loops that would explain Parkinson's were a local "vicious circle" – namely a closed loop involving only the spine and the muscles but not the brain – or the effect of a closed loop of activity in the central nervous system, which sent a cyclical signal to the region of the body affected by the tremor. McCulloch and Wortis wondered whether other diseases, such as epilepsy, could be explained by closed loops of neural activity. They did not consider that closed loops of activity could be a normal feature of the nervous system, in part because their discussions were taking place before Lawrence Kubie published the first theoretical paper postulating closed loops in the central nervous system to explain memory (Kubie 1930).[12] Later in his life, McCulloch would hypothesize closed loops as explanations for many normal neural functions.

In 1929, McCulloch had a new insight. It occurred to him that the all-or-none electric impulses transmitted by each neuron to its neighbors might correspond to the mental atoms of his psychological theory, where the relations of excitation and inhibition between neurons would perform logical operations upon electrical signals corresponding to inferences of his propositional calculus of psychons. His psychological theory of mental atoms turned into a theory of "information flowing through ranks of neurons."[13]

This was McCulloch's first attempt "to apply Boolean algebra to the behavior of nervous nets."[14] The brain would embody a logical calculus like that of Whitehead and Russell's *Principia Mathematica*, which would account for how humans could perceive objects on the basis of sensory signals and how humans could do mathematics and abstract thinking. This was the beginning of McCulloch's search for the "logic of the nervous system," on which he kept working until his death. A major difficulty for the formulation of his logical calculus was the treatment of closed loops of neural activity. McCulloch was trying to describe the causal structure of neural events by assigning temporal indices to them. But he thought a closed loop meant that an event could be its own ancestor, which did not make sense to him. He wanted "to close the loop" between chains of neuronal events – that is, to find a mathematical representation of the relationship between events in a closed loop – but did not know how to relate the events in the closed loops to one another. He would not find a solution to this difficulty until he met Walter Pitts in the early 1940s.[15]

In 1934, McCulloch moved to Yale to work in Joannes Dusser de Barenne's Laboratory of Neurophysiology. Dusser de Barenne was a distinguished Dutch neurophysiologist who had moved from Holland to Yale in 1930.[16] With Dusser de Barenne, McCulloch worked mostly on mapping the connections between brain areas. McCulloch worked at Yale until shortly after Dusser de Barenne's death in 1940. McCulloch's work during those years launched his academic career.[17]

At Yale, McCulloch also attended a philosophical seminar for research scientists organized by Filmer Northrop, a member of the Philosophy Department. At one of those seminars, Frederic Fitch, a distinguished logician from the same department, presented the theory of deduction from the *Principia Mathematica*. McCulloch also attended advanced lectures by Fitch on logical operators and urged Fitch to work on the logic of neural nets.[18]

While McCulloch was at Yale, he became acquainted with the work of Woodger (1937), who advocated the axiomatic method in biology. In a letter to a colleague written in 1943, McCulloch wrote:

> I personally became acquainted with Woodger because the great interest of the biologists in Yale had led to his coming thither to tackle some of their problems. When he finally departed, it was not because they were not convinced of the value of his attempt but because he was convinced that the ambiguity of their statements prevented logical formulation. It was to discussions with him and with Fitch that I owe much of my persistence in attempting a logical formulation of neuronal activity. Until that time I had merely used the nomenclature of the *Principia Mathematica* to keep track of the activity of neuronal nets.[19]

In the same letter, McCulloch suggested that it was only around this time that he started seeing his theory of the brain as a "theory of knowledge":

> [T]he theory... began originally as a mere calculus for keeping track of observed realities. It was at work for seven years before it dawned on me that it had those logical implications which became apparent when one introduces them into the grandest of all feed-back systems, which runs from the scientist by manipulations through the objects of this world, back to the scientist – so producing in him what we call theories and in the great world are little artifacts.[20]

McCulloch's theory had come a long way. In the mid-1920s, it began as a theory of mental atoms and their associations. By the end of the 1920s, it identified mental atoms with neuronal pulses and used logic to represent relations between neural realizations of mental atoms. Seven years later, McCulloch turned the "logical" relations between neural realizations of mental atoms into an explanation of how scientific theories are constructed, and more generally of how humans can gain knowledge.

McCulloch had known Northrop since 1923, and the two of them kept in touch. Northrop wrote extensively on science and scientific methodology. He believed that scientific disciplines reach maturity when they start employing logic and mathematics in formulating rigorous, axiomatic theories:

> The history of science shows that any empirical science in its normal healthy development begins with a more purely inductive emphasis, in which the empirical data of its subject matter are systematically gathered, and then comes to maturity with deductively-formulated theory in which formal logic and mathematics play a most significant part.[21]

Northrop argued that biology was finally reaching its maturity with the work of Woodger (1937) and Rashevsky (1938), a Russian

physicist who had imported formalisms and techniques from mathematical physics into biology.[22]

In 1936, Alan Turing published his famous paper on computability (Turing 1936–37), in which he introduced Turing Machines and used them to draw a clear and rigorous connection between computing, logic, and machinery. In particular, Turing argued that any effectively calculable function can be computed by some Turing Machine – a thesis now known as the Church–Turing thesis (CT) – and proved that some special Turing Machines, which he called "universal," can compute any function computable by Turing Machines.[23] By the early 1940s, McCulloch had read Turing's paper. In 1948, in a public discussion during the Hixon Symposium, McCulloch declared that in formulating his theory of mind in terms of neural mechanisms, reading Turing's paper led him in the "right direction."[24]

2. MATHEMATICAL BIOPHYSICS OF THE CENTRAL NERVOUS SYSTEM

In September 1941, McCulloch moved to the University of Illinois in Chicago, where he made contact with the Committee on Mathematical Biology. The Committee, a pioneering research group in biophysics at the University of Chicago, was a creation of Nicolas Rashevsky.

Rashevsky advocated the development of mathematical models of idealized *biological* processes, analogously to how theoretical physicists develop mathematical models of idealized *physical* processes.[25] His goal was a "systematic mathematical biology, similar in aim and structure to mathematical physics"[26]; "mathematical biology would stand in the same relation to experimental biology as mathematical physics stands to experimental physics."[27] Rashevsky's method involved making assumptions about the essential features of a biological process, representing those features as mathematical variables, connecting the variables through equations that satisfy the initial assumptions, and studying the equations – especially differential and integral equations – to investigate possible mechanisms underlying the biological process under investigation and make testable predictions about it. For reasons of mathematical tractability, the models so constructed – at least initially – had to be simplified and idealized relative to real biological processes.

In several writings on methodology, Rashevsky extensively addressed the objection that his mathematical approach was irrelevant

to real biological processes. He considered two reasons for irrelevance: the "necessary abstraction and oversimplification that is inherent to this approach" and "the frequent discussion of some purely theoretical cases which have no immediate 'practical' interest."[28] Rashevsky responded that his method was precisely the one that had proved fruitful in physics. Physicists started with simplistic models – such as elastic billiard balls as models for gas molecules – and then progressively refined those simplistic models into more realistic ones. Furthermore, Rashevsky continued,

> A theoretical problem may have an interest of its own, and should not be tabooed only because *at present* it does not appear applicable to a definite experiment. The history of physics shows how frequently such 'purely theoretical' developments led, a few decades later, to the most astonishing practical results.[29]

We shall see that in formulating their theory, McCulloch and Pitts followed Rashevsky's precepts quite closely. The only difference is that instead of the continuous mathematics of differential and integral equations, they used the discrete mathematics of logic and computation theory.

Since cells are the building blocks of organisms, Rashevsky began his efforts by formulating a theory of cell metabolism. His theory included a study of metabolic processes that might be responsible for cell division.[30] His next research project involved the study of excitation and conduction of impulses in nerve fibers.[31] A third major project concerned the mathematical biophysics of the central nervous systems. The third project's goal was the explanation of psychological phenomena in terms of systems of interconnected nerve fibers.[32]

Rashevsky's theory of the nervous system was based on the all-or-none law of nervous activity, according to which neurons either pulse or remain at rest depending on whether their input is above a certain threshold. As a consequence of the all-or-none law, Rashevsky followed mainstream physiologists in assuming that the physiological variable that was relevant to nervous activity was the frequency of impulses traveling along nerve fibers. Furthermore, Rashevsky accepted the popular view that there were both excitatory and inhibitory connections among fibers. Accordingly, he defined two quantities, ε and ι, which represented excitatory and inhibitory effects of afferent fibers, and built his theory on the assumption that those two quantities determined the excitation level, and thus the frequency of impulses, of efferent fibers. Rashevsky studied several patterns of activity that resulted from different combinations of ε and ι in

combination with different "geometrical arrangement of the neuro-elements."[33] He also applied his theory by offering possible mechanisms for the phenomena of conditioned reflexes, "discrimination of relations" such as "larger than" and "smaller than," Gestalt effects, and even "rational learning and thinking."[34]

In the late 1930s, Rashevsky formed and led a group of talented collaborators, who helped him develop his theoretical framework and apply it to new areas. Rashevsky's most important associates on the biophysics of the central nervous system were Alston Householder and Herbert Landahl. With them and others, Rashevsky applied his theory to explain various psychological phenomena, including "discrimination of intensities," "psychophysical discrimination," and "perception of visual patterns."[35]

Among Householder's many contributions to Rashevsky's program was a "theory of steady-state activity in nerve-fiber networks."[36] Householder noted that in general, "the behavior of any complex of nerve fibers must depend alone upon the dynamic properties of the individual fibers and the structural relations among these fibers. "[37] "Hence," he concluded, "there must necessarily exist certain general propositions which express the activity of any nerve-fiber complex in terms of the dynamics of the individual fibers and of the structural constants."[38] Householder began his search for those general propositions by considering the steady-state activity of nerve-fiber networks under constant stimulation. He assumed that upon constant stimulation, a nerve fiber reaches a steady-state activity whose intensity is a linear function of the stimulus. He then investigated structures of nerve-fiber networks made out of "circuits," that is, closed loops of fibers. (These were the same closed loops that McCulloch was concerned with, but there is no evidence that McCulloch studied Householder's work.) Given a nerve-fiber network with a certain structure, Householder studied the patterns of steady-state activity that the network could exhibit as a function of its stimuli. Householder's theory of steady-state activity in nerve-fiber networks was the main starting point for Walter Pitts's work on the biophysics of the central nervous system.

Pitts told some of his friends that when he was about 12, he found a copy of the *Principia Mathematica* in a public library, and studied it cover-to-cover.[39] Around 1938, at the age of 15, Pitts fled his parental home and never spoke to his family again. Around the same time, he attended a lecture by Bertrand Russell at the University of Chicago. During the lecture, he met an 18-year-old member of the audience,

Jerome Lettvin, who was preparing for medical school by studying biology at the University of Chicago. Pitts and Lettvin became best friends.[40] In the late 1930s, Pitts attended classes at the University of Chicago, but without enrolling as a student. He studied logic with Rudolf Carnap and biophysics with Rashevsky. Pitts became a member of Rashevsky's group, and soon began to produce original research without ever earning a degree.

According to Pitts, Householder "suggested the problem" on which Pitts wrote three of his first papers; Pitts also expressed "appreciation" for Householder's "perspicacious counsel and criticisms."[41] In those three papers, Pitts generalized Householder's theory into what he called a "theory of neuron networks," which extended and unified Householder's results in a simpler way.[42] To achieve this, Pitts simplified his mathematical treatment by making an important assumption. He defined the

> ... total conduction time of a fiber as the sum of its conduction time and the synaptic delay at the postliminary synapse: we shall suppose ... that all the total conduction times of fibers of the circuit *C* in question are equal: and we shall measure time so that this quantity is unity.[43]

Synaptic delay was the time between the arrival of a neuron's inputs and the beginning of its own output. Pitts assumed that all fibers in a network were active during temporal intervals of equal duration, and that the events occurring during one interval only affected the events occurring during the following interval. An analogous assumption would later play an important role in McCulloch and Pitts's theory.

Like Householder's theory, Pitts's theory covered "circuits," i.e. networks with closed loops of activity. Unlike Householder's theory, Pitts's theory covered not only the networks' steady-state activity but also their "non-steady-state activity and the conditions under which a steady-state may be attained."[44] Pitts distinguished between the "static problem," namely, that of finding equilibrium points at which a network exhibits steady-state activity, and the "dynamic problem," namely, that of determining whether the equilibrium points are stable or unstable. Pitts solved both the static and the dynamic problems for the networks defined by his theory. He also formulated the "inverse network problem," that is, "given a preassigned pattern of activity over time, to construct when possible a neuron-network having this pattern."[45] Pitts solved a special case of the inverse network problem.

Pitts's work was not limited to his theory of neuron networks. He developed a "general theory of learning and conditioning."[46] In this case, Pitts's theory was not formulated in terms of neural mechanisms, but it characterized the relation between stimuli and responses mathematically, aiming to capture "the contribution of each trial to learning in a way depending upon its relevant characteristics and those of the previous trials."[47] Around the same time, Pitts also collaborated on an elaborate prank with Lettvin, his medical student friend. Following Rashevsky's methodology, they chose two variables, φ and ψ, to represent the intensity of emotion and the intensity of activity of a subject. They then wrote differential equations representing the variation and mutual interaction of φ and ψ over time. They solved the equations and presented an analysis of the stability and instability of the system's equilibrium points. Lettvin and Pitts's theory was published in Rashevsky's *Bulletin of Mathematical Biophysics* as a "mathematical theory of affective psychoses," purportedly covering "circular insanities," "reactive psychoses," and the "catatonia of Kahlbaum."[48] Their paper contained no explicit indication of its satirical character and according to Lettvin, no one got their joke.[49]

According to Lettvin, it was during his years with the Rashevsky group – before meeting McCulloch – that Pitts developed the view that the brain is a "logical machine."[50] Using anachronistic terminology, Lettvin put it as follows:

> 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.[51]

Unfortunately, there is no independent evidence that Lettvin's recollection here is correct.[52] In the papers that Pitts wrote independently of McCulloch, Pitts did not suggest that the brain is a logic machine. Before McCulloch entered the picture, neither Pitts nor any other member of Rashevsky's biophysics group employed logical or computational language to describe the functions performed by networks of neurons. The use of logic and computation theory to model the brain and understand its function appeared for the first time in McCulloch and Pitts's 1943 paper; this is likely to be a contribution made by McCulloch to his joint project with Pitts.

Soon after McCulloch met Pitts, around the end of 1941, they started collaborating on a joint mathematical theory that employed

logic to model nervous activity, and they worked on it during the following two years. They worked so closely that Pitts (as well as Lettvin) moved in with McCulloch and his family for about a year in Chicago. McCulloch and Pitts became intimate friends and they remained so until their death in 1969.[53] According to McCulloch, they worked largely on how to treat closed loops of activity mathematically, and the solution was worked out mostly by Pitts using techniques that McCulloch didn't understand. To build up their formal theory, they adapted Carnap's rigorous (but cumbersome) formalism, which Pitts knew from having studied with Carnap. Thus, according to McCulloch, Pitts did all the difficult technical work.[54] The resulting paper was published in the *Bulletin of Mathematical Biophysics* in December 1943, with a brief follow-up written by McCulloch and Pitts with Herbert Landahl, in which they attempted to connect their theory to Rashevsky's theory of nervous activity by proposing a way to define Rashevsky's quantities ε and ι in terms of their theory.[55]

3. SOLVING THE MIND–BODY PROBLEM

A formidable obstacle to a theory of mind based on neural mechanisms was the mind–body problem. In a commentary to a paper presented in May 1943 at the Illinois Psychiatric Society, McCulloch explained:

> We have a dichotomy in medicine, which has grown increasingly... Psychiatric approach on one side, particularly the psychoanalytic approach, has produced one group; the organic approach to the physiology of particular organs and disease processes has made organicists of another group. It has grown difficult for us to talk to each other. I am afraid that there is still in the minds of most of us, and that there probably will be for years, that difficulty which concerned and still concerns many thinking people – I mean the dichotomy between mind and body.[56]

There were "two types of terminology," McCulloch continued: "mental terms" described "psychological processes, for these exhibit ideas and intentions," whereas "physical terms" described "bodily processes, for these exhibit matter and energy." But:

> ... it remains our great difficulty that we have not ever managed to conceive how our patient – our monad – can have a psychological aspect and a physiological aspect so divorced. You may think that I am exaggerating the difficulty here, but there have appeared within the last few years two books which tilt at the same windmill. One is Sherrington, called "Man and His Nature," and in it Sherrington, the marvelously

honest physiologist, attempts to make head and tail of the mind–body relation, but is frustrated because in that world "Mind goes more ghostly than a ghost." The other book, by Wolfgang Koehler (the founder of Gestalt psychology), is entitled "The Place of Value in a World of Fact," but in spite of his endless searching, you will be convinced that he has not found the place of value in the world of fact. Such was the unsatisfactory state of our theory until very recently.[57]

After thus stating the mind–body problem, McCulloch pointed at two recent developments that gave hope for its solution.

As an answer to the question of "the place of values in a world of fact," McCulloch cited the newly published work of Rosenblueth, Wiener, and Bigelow (1943), which used the notion of feedback to account for teleological behavior. As to what McCulloch called the "formal" aspect of mind, he promised he was going to have something to contribute soon:

At the present time the other mental aspect of behavior – I mean its ideational or rational, formal or logical aspect – is coming to the fore. This work ... should be coming to fruition in the next year or two... We do resent the existing hiatus between our mental terminology and our physical terminology. It is being attacked in a very realistic fashion today. So while we do at the moment think of it as a "leap from psyche to soma," we are busy bridging the gap between mental processes and physical processes. To this audience it is interesting that that bridge is being made by demonstrating that the properties of systems which are like our nervous system necessarily show those aspects of behavior that make us call it "mental" – namely, ideas and purposes.[58]

In the last sentence, McCulloch was suggesting that the formal equivalence between chains of neural events and certain logical inferences, which he saw as relations between "ideas," "necessarily" turned the former into events with "mental" properties, like the latter. This slide from formal properties of neural events to mental properties, in which the former are taken to be sufficient ground for the latter, will strike most readers as fallacious. Nonetheless, it is a prominent characteristic of McCulloch's thinking about brains.[59] The explanation for the "formal" aspect of the mind, and hence the solution to that component of the mind–body problem, was about to be offered by McCulloch in the paper he was writing with Pitts. Their way of solving the problem was to "demonstrate" how a system of neuron-like elements embodies ideas by having a causal structure that mirrors propositional inferences.

In a letter written a few months before the commentary cited above, McCulloch was more detailed and explicit as to what he hoped to accomplish with his theory and the role that logic played in it:

As to the "formal" properties [of the mind], it is perfectly possible today (basing the work on the all-or-none law and the requirement of summation at a synapse and of inhibition either at a synapse or by preoccupation of a requisite pool of internuncials) to show that neuronal reactions are related to antecedent neuronal reactions – I mean reactions in parts of the nervous system afferent to the reaction in question – in a manner best schematized by symbolic logic; in brief, that the efferent impulses are related to the afferent impulses as logical consequences are related to logical antecedents, and hence that classes of the latter are so related to classes of the former.

Little consideration is necessary to show that neuronal and all other reactions which derive their energy metabolically and are triggered off by something else, being reactions of the zero order with respect to what initiates them, bear to their precipitating causes the same relation that propositions do to that which they propose. If then, from the sense organ forward, the reaction of subsequent neurones is dependent upon any selection from the totality of energy delivered to the system, the response corresponds to an abstraction from that totality, so that neural behavior is not only essentially propositional but abstract with respect to its precipitating cause.[60]

Once again, McCulloch was describing the work he was pursuing with Pitts. The all-or-none law allowed McCulloch and Pitts to use symbolic logic to describe neural activity, so that inferential relations among propositions described causal streams of neural events. This, for McCulloch, was enough to show that "neural behavior is essentially propositional" in a way that explained mechanistically the "formal" aspect of the mind.

The sense in which neural behavior was essentially propositional was further clarified by McCulloch in a letter to a neurophysiologist at the University of Chicago, Ralph Lillie. In February 1943, he explained how "we might be able to see mechanistically the problem of ideas":

[W]hat was in my mind was this: that neuronal activity bore to the world external to the organism the relationship that a proposition bears to that to which it proposes. In this sense, neuronal activity so reflects the external world as to account for that all-or-none characteristic of our logic (and of our knowledge) which has been one of the greatest stumbling blocks to epistemology. I think that for the first time we are in a position to regard scientific theory as the natural consequence of the neuronal activity of an organism (here the scientist) ... And this has come about because the observed regularity – all-or-none of neurones, bears a one-to-one correspondence to those peculiar hypothetical psychic atoms called psychons which preserve in the unity of their occurrence both the all-or-none law and the property of reference characteristic of propositions.[61]

Thanks to the all-or-none law, neural pulses stood in "one-to-one correspondence" to psychons, and just like psychons and propositions, neuronal activity had "the property of reference."[62]

We have seen that McCulloch had a theory of mental atoms, the psychons, which he identified with all-or-none neural events. In today's terms, McCulloch was convinced that neural pulses had contents corresponding to the contents of mental atoms, and that human knowledge could be explained by the logical relations between those contents. With these ideas, McCulloch wanted to reduce the mind to the brain, explain how the brain acquires knowledge, and thus solve the mind–body problem. The solution, he thought, lied in a theory of the brain that employed logic and computation theory to show how the brain draws inferences and represents numbers. But McCulloch was not enough of a mathematician to formulate the theory by himself. Pitts had the technical prowess that McCulloch lacked. Pitts was familiar with Rashevsky's methodology of building mathematical theories of idealized biological systems. Pitts had already produced his own "theory of neuron networks," which covered those closed loops of neural activity that McCulloch considered crucial to many neural and psychological phenomena. But Pitts's theory did not employ logic or computation theory, and because of this, it did not suit McCulloch's purposes. McCulloch and Pitts joined forces and produced a brand new theory.

They called their product "a logical calculus of the ideas immanent in nervous activity." As logician Frederic Fitch pointed out in reviewing their paper for the *Journal of Symbolic Logic*, this was not quite a logical calculus in the sense employed by logicians.[63] Strictly speaking, a calculus is a combination of a grammar and a deductive system. A grammar is a set of symbols and rules for determining which sequences of symbols constitute well-formed expressions. A deductive system specifies which (if any) expressions can be taken as axioms and which output expressions can be derived from input expressions. There must also be an effective method for determining whether a chain of expressions satisfies the rules of the deductive system. A *logical* calculus is a calculus whose deductive system is aimed at capturing *logical* inferences among the class of expressions allowed by the calculus' grammar.

As we shall see, McCulloch and Pitts defined a grammar for representing a class of expressions, but did not define a deductive system. Instead, they made a number of assumptions about the properties of neurons and their pulses, so as to define a class of nets of idealized neurons. They also interpreted the expressions defined by their grammar as describing input–output behaviors of neurons. Finally, they devised effective methods for mapping classes of

expressions and nets one-to-one, in such a way that each neuron's behavior was correctly described by one of their well-formed expression, and each well-formed expression was satisfied by some neuron's behavior. If the neuronal pulses were interpreted, in turn, as "propositions" or "ideas," then the expressions describing neuronal behavior could be seen as describing logical relations between a neuron's inputs and its outputs.

4. MOTIVATION

McCulloch and Pitts's paper started by rehearsing some established neurophysiological facts: the nervous system is a system of neurons connected through synapses, neurons send excitatory and inhibitory pulses to each other,[64] and each neuron has a threshold determining how many excitatory and inhibitory inputs are necessary and sufficient to excite it at a given time.[65]

Then, the authors introduced the main premise of their theory: that neuronal signals are "equivalent" to propositions. This was presumably what justified their title, which mentions *ideas* immanent in nervous activity. They introduced this theme in a curious and oblique way, appealing not to some explicit motivation but to "considerations," made by one of the authors, which they did not give:

> Many years ago one of us, by *considerations impertinent to this argument*, was led to conceive of the response of any neuron as *factually equivalent* to a proposition which proposed its adequate stimulus. He therefore attempted to record the behavior of complicated nets in the notation of the symbolic logic of propositions. 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. To each reaction of any neuron there is a corresponding assertion of a simple proposition. This, in turn, implies either some other simple proposition or the disjunction or the conjunction, with or without negation, of similar propositions, according to the configuration of the synapses upon and the threshold of the neuron in question.[66]

In light of the previous sections, the author of the "considerations" was McCulloch, and the considerations were those that led him to formulate first his theory of psychons, and then his theory of information flow through ranks of neurons. A proposition that "proposes a neuron's adequate stimulus" was a proposition to the effect that the

neuron receives a certain input at a certain time. The authors did not explain what they meant by "factual equivalence" between neuronal pulses and propositions, but their language, the background discussed above, and the conclusions they drew from their theory (see Section 8 below), suggests they meant both that neuronal pulses are represented by propositions, and that neuronal pulses have propositional content.

The theory was divided into two parts: one part for nets *without* "circles" (McCulloch and Pitts's term for closed loops of activity); the other for nets *with* circles or "*cyclic* nets." (Today cyclic nets are called "recurrent networks.") The authors pointed out that the nervous system contains many circular, "regenerative" paths.[67] The term "circle" may have been borrowed from Turing (1936–37), who had used "machines with circles" for Turing Machines whose computations continue forever without producing the desired output, and "machines without circles" for Turing Machines that produce the desired output. Like a Turing Machine with circles, a net with circles may run forever if left unperturbed.

5. ASSUMPTIONS

In formulating their theory, McCulloch and Pitts made the following five assumptions:

1. The activity of the neuron is an "all-or-none" process.
2. A certain fixed number of synapses must be excited within the period of latent addition in order to excite a neuron at any time, and this number is independent of previous activity and position of the neuron.
3. The only significant delay within the nervous system is synaptic delay.
4. The activity of any inhibitory synapse absolutely prevents excitation of the neuron at that time.
5. The structure of the net does not change with time.[68]

These assumptions idealize the then known properties of neurons. Assumption (1) is simply the all-or-none law: neurons were believed to either pulse or be at rest. As to (4) and (5), McCulloch and Pitts admitted that they are false of the nervous system. Under the other assumptions, however, they argued that nets that do not satisfy (4) and (5) are functionally equivalent to nets that do.[69]

As to (2) – that a fixed number of neural stimuli is always necessary and sufficient to generate a neuron's pulse – it is a radical simplification. First, as McCulloch and Pitts discussed in their preliminary review of neurophysiology, the "excitability" of a neuron does vary over time: after a neuron's pulse, there is an absolute refractory period in which the neuron cannot pulse at all, and "[t]hereafter its excitability returns rapidly, in some cases reaching a value above normal from which it sinks again to a subnormal value, whence it returns slowly to normal."[70] Second, as McCulloch and Pitts also discussed, synapses are plastic, so that which stimuli will activate a neuron is a function of the previous patterns of activity and stimulation of that neuron. Some changes in the responsiveness of neurons to stimuli are temporary, as in "facilitation and extinction," whereas other changes are permanent and constitute "learning."[71] With respect to this second point, however, McCulloch and Pitts argued that a net whose responsiveness to stimuli varies can be replaced by a formally equivalent net whose responsiveness is fixed.[72] Hence, synaptic plasticity does not affect their conclusion that neural events satisfy the "logic of propositions."[73]

Assumption (3) was crucial. As Pitts had done in his previous theory of neuron networks, McCulloch and Pitts assumed that all neurons within a network were synchronized so that all the relevant events in the network – conduction of the impulses along nerve fibers, refractory periods, and synaptic delays – occurred within temporal intervals of fixed and uniform length. Moreover, they assumed that all the events within one temporal interval only affected the relevant events within the following temporal interval. Assumption (3) had the effect of discretizing the dynamics of the net: instead of employing differential and integral equations to describe frequencies of impulses, as the members of Rashevsky's group did, McCulloch and Pitts could now describe patterns of individual neural impulses as a discrete function of previous patterns of neural impulses. Logical functions of discrete states could be used to fully describe the transitions between neural events. In pursuit of McCulloch's dream of a theory of knowledge based on logic, McCulloch and Pitts could now replace the continuous mathematics of the physicists – the mathematics used by Rashevsky, and by Pitts before meeting McCulloch – with the discrete mathematics of the logicians as the appropriate tool for modeling the brain and studying its functions.

McCulloch and Pitts were perfectly aware that the neuron-like elements in their theory were distant from real neurons: "Formal

neurons were deliberately as impoverished as possible.''[74] Idealization and simplification were, after all, basic aspects of the biophysical methodology advocated by Rashevsky. In a letter written to a colleague asking for clarification after a public presentation of the theory, McCulloch further explained that their theory ignored any neural malfunction:

> [W]e in our description restricted ourselves to the regular behavior of the nervous system, knowing full well that irregularities can be and are frequently brought about by physical and chemical alterations of the nervous system. As a psychiatrist, I am perhaps more interested in these than in its regular activity, but they lead rather to a theory of error than a theory of knowledge, and hence were systematically excluded from the description.[75]

In McCulloch's eyes, the differences between real neurons and the elements employed in his theory were inessential. His goal was not to understand neural mechanisms *per se*, but rather to explain how something close enough to a neural mechanism could – in McCulloch's words – exhibit ''knowledge,'' the kind of ''ideational,'' ''rational,'' ''formal,'' or ''logical'' aspect that McCulloch associated with the mind (cf. Section 3 above). McCulloch's goal was to offer an explanation of the mind in terms of neural-like mechanisms. Since McCulloch thought the explanation had to involve logic to describe neural activity, what was needed was a set of simplifications and idealizations that allowed logic to be so used.

6. NETS WITHOUT CIRCLES

McCulloch and Pitts's technical language was cumbersome; here their theory is given in a slightly streamlined form that makes it easier to follow. The neurons of a net N are denoted by $c_1, c_2, \ldots c_n$. A primitive expression of the form $N_i(t)$ means that neuron c_i fires at time t. Expressions of the form $N_i(t)$ can be combined by means of logical connectives to form complex expressions that describe the behavior of different neurons at certain times. For example, $N_1(t)$ & $N_2(t)$ means that neurons c_1 and c_2 fire at time t, $N_1(t-1) \vee N_2(t-2)$ means that either c_1 fires at $t-1$ or c_2 fires at $t-2$ (or both), etc. These complex expressions can in turn be combined by the same logical connectives. As well formed combinations, McCulloch and Pitts allowed only the use of conjunction (A & B), disjunction ($A \vee$B), conjunction and negation ($A \& \sim B$), and a special connective S that shifts the temporal index of an expression backwards in time, so that

$S[N_i(t)] = N_i(t-1)$. A complex expression formed from a number of primitive expressions $N_1(t), \ldots N_n(t)$ by means of the above connectives is denoted by $Expression_j[N_1(t), \ldots N_n(t)]$. In any net without circles, there are some neurons that receive no inputs from other neurons; these are called *afferent* neurons.

The two main technical problems McCulloch and Pitts formulated and solved were "to calculate the behavior of any net, and to find a net which will behave in a specified way, when such a net exists."[76] These problems were analogous to those formulated by Householder and Pitts within their earlier theories. Householder had asked which general propositions "express the activity of any nerve-fiber complex in terms of the dynamics of the individual fibers and of the structural constants."[77] Pitts had added the inverse network problem of, "given a preassigned pattern of activity over time, to construct when possible a neuron-network having this pattern."[78] Within McCulloch and Pitts's theory, however, a net's pattern of activity was no longer described as a frequency of impulses, but rather as a precise pattern of neuronal impulses described by a logical formula.

In terms of McCulloch and Pitts's theory, the two problems can be formulated as follows:

First problem: given a net, find a class of expressions $\boldsymbol{C}$ such that for every neuron c_i, in $\boldsymbol{C}$ there is a true expression of the form

$$N_i(t) \text{ if and only if Expression}_j [(N_{i-g}(t-1), \ldots N_{i-2}(t-1), N_{i-1}(t-1)],$$

where neurons $c_{i-g}, \ldots c_{i-2}$, and c_{i-1} have axons inputting c_i.

The significance of this expression is that it describes the behavior of any (non-afferent) neuron in terms of the behavior of the neurons that are afferent to it. If a class $\boldsymbol{C}$ of such expressions is found, then propositional logic can describe the behavior of any non-afferent neuron in the net in terms of the behavior of the neurons afferent to it.

Second problem: given an expression of the form

$$N_i(t) \text{ if and only if Expression}_j [(N_{i-g}(t-1), \ldots N_{i-2}(t-1), N_{i-1}(t-1)],$$

find a net for which it is true.

McCulloch and Pitts showed that in the case of nets without circles, these problems were easily solved. To solve the first problem,

they showed how to write an expression describing the relation between the firing of any neuron in a net and the inputs it received from its afferent neurons. To solve the second problem, they showed how to construct nets that satisfy their four combinatorial schemes (conjunction, disjunction, conjunction-cum-negation, and temporal predecessor), giving diagrams that showed the connections between neurons that satisfy each scheme (Figure 1). Then, by induction on the size of the nets, all expressions formed by those combinatorial schemes are satisfiable by McCulloch–Pitts nets.[79]

By giving diagrams of nets that satisfy simple logical relations between propositions and by showing how to combine them to satisfy more complex logical relations, McCulloch and Pitts developed a powerful technique for designing circuits that satisfy given logical functions by using a few primitive building blocks. This is the main aspect of their theory used by von Neumann in describing the design of digital computers.[80] Today, McCulloch and Pitts's technique is part of logic design, an important area of computer design devoted to digital circuits for computers. The building blocks of contemporary logic design are called logic gates. In today's terminology, McCulloch and Pitts's nets are logic gates and combinations of logic gates.[81]

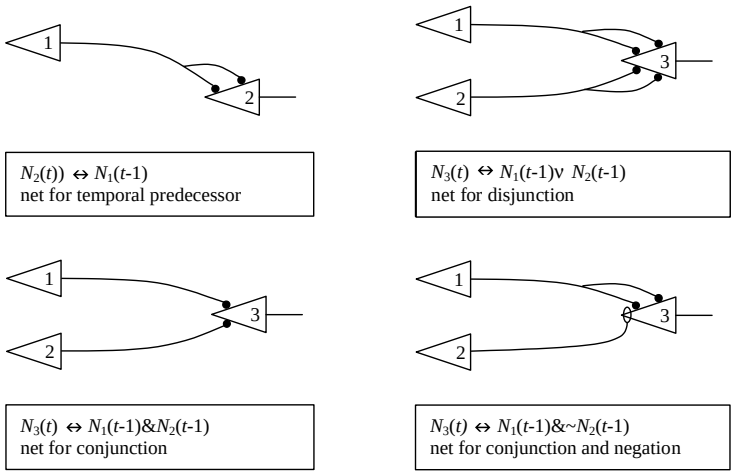

Figure 1. Diagrams of McCulloch and Pitts nets. In order to send an output pulse, each neuron must receive two excitory inputs and no inhibitory inputs. Lines ending in a dot represent excitatory connections; lines ending in a hoop represent inhibitory connections.

The original purpose of McCulloch and Pitts's technique for designing nets was to explain mental phenomena. As an example, they offered an explanation of a well-known heat illusion by constructing an appropriate net. A cold object touching the skin normally causes a sensation of cold, but if it is held for a very brief time and then removed, it can cause a sensation of heat. In designing their net, McCulloch and Pitts reasoned as follows. They started from the known physiological fact that there are different kinds of receptors affected by heat and cold, and they assumed that there are neurons whose activity "implies a sensation" of heat.[82] Then, they assigned one neuron to each function: heat reception, cold reception, heat sensation, and cold sensation. Finally, they observed that the heat illusion corresponded to the following relations between three neurons: the heat-sensation neuron fires either in response to the heat receptor or to a brief activity of the cold receptor (Figure 2).

McCulloch and Pitts used this example for a general observation about the relation between perception and the world:

> This illusion makes very clear the dependence of the correspondence between perception and the "external world" upon the specific structural properties of the intervening nervous net.[83]

Then they pointed out that, under other assumptions about the behavior of the heat and cold receptors, the same illusion could be

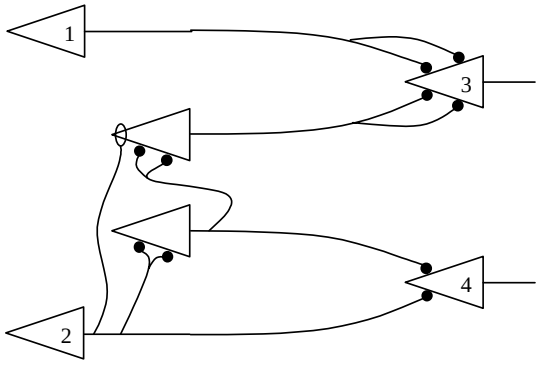

Figure 2. Net explaining the heat illusion. Neuron 3 (heat sensation) fires if and only if it receives two inputs, represented by the lines terminating on its body. This happens when either neuron 1 (heat reception) fires or neuron 2 (cold reception) fires once and then immediately stops firing. When neuron 2 fires twice in a row, the intermediate (unnumbered) neurons excite neuron 4 rather than neuron 3, generating a sensation of cold.

explained by different nets.[84] They were not proposing their network as the actual mechanism behind the heat illusion, but rather as a possible mechanism that explained the illusion.

7. NETS WITH CIRCLES, COMPUTATION, AND CT

The problems for nets with circles were analogous to those for nets without circles: given the behavior of a neuron's afferents, find a description of the behavior of the neuron; and find the class of expressions and a method of construction such that for any expression in the class, a net could be constructed that satisfies the expression. The authors pointed out that the theory of nets with circles is more difficult than the theory of nets without circles. This is because activity around a circle of neurons may continue for an indefinite amount of time, hence expressions of the form $N_i(t)$ may have to refer to times that are indefinitely remote in the past. For this reason, the expressions describing nets with circles are more complicated, involving quantification over times. McCulloch and Pitts offered solutions to the problems of nets with circles, but their treatment of this part of the theory was obscure, admittedly sketchy,[85] and contained some errors. As a consequence, it is almost impenetrable.[86]

At the end of this section, McCulloch and Pitts drew the connection between their nets and computation:

> It is easily shown: first, that every net, if furnished with a tape, scanners connected to afferents, and suitable efferents to perform the necessary motor-operations, can compute only such numbers as can a Turing machine; second, that each of the latter numbers can be computed by such a net; and that nets with circles can be computed by such a net; and that nets with circles can compute, without scanners and a tape, some of the numbers the machine can, but no others, and not all of them. This is of interest as affording a psychological justification of the Turing definition of computability and its equivalents, Church's λ-definability and Kleene's primitive recursiveness: If any number can be computed by an organism, it is computable by these definitions, and conversely.[87]

This brief passage is the only one mentioning computation. By stating that McCulloch–Pitts nets computed, this passage provided the first known published link between the mathematical theory of computation and brain theory. It was a pivotal statement in the history of computationalism.

It is often said that McCulloch and Pitts proved that their nets can compute anything that Turing Machines can compute (e.g., Koch and Segev 2000). This misconception was propagated by McCulloch

himself. For instance, in summarizing the significance of their paper, McCulloch wrote to a colleague:

> [T]he original paper with Pitts entitled "A Logical Calculus of Ideas Immanent in Nervous Activity" ... sets up a calculus of propositions subscripted for the time of their appearance for any net handling all-or-none signals, and shows that such nets can compute any computable number or, for that matter, do anything any other net can do by the way of pulling consequences out of premises.[88]

But in discussing computation in their paper, McCulloch and Pitts did not *prove* any results about the computation power of their nets; they only stated that there were results to prove. And their conjecture was not that their nets can compute anything that can be computed by Turing Machines. Rather, they claimed that *if* their nets were provided with a tape, scanners, and "efferents," *then* they would compute what Turing Machines could compute; without a tape, McCulloch and Pitts expected even nets with circles to compute a smaller class of functions than the class computable by Turing Machines.

Their comment that these conjectures were "easily shown" may suggest that the proofs were trivial. On the contrary, the question of what was computable by McCulloch–Pitts nets was not even explicitly defined by the authors. Several years later, Stephen Kleene set out to rigorously formulate and solve the problem of what McCulloch–Pitts nets can compute. Kleene proceeded independently of McCulloch and Pitts's treatment of nets with circles because he found it "obscure" and because he found an "apparent counterexample."[89] Kleene defined the notion of "regular events" (today called "regular languages") and proved that McCulloch–Pitts nets can "represent" regular events (in today's terminology, they can accept regular languages). In the same paper, Kleene also defined an alternative formalism, which generalized McCulloch–Pitts nets by allowing the "cells" in the network to take any of a finite number of internal states. Kleene called his new formalism "finite automata," and showed that McCulloch–Pitts nets are computationally equivalent to finite automata.

McCulloch and Pitts did not explain what they meant by saying that nets compute. As far as the first part of the passage is concerned, the sense in which nets compute seems to be a matter of *describing* the behavior of nets by the vocabulary and formalisms of computability theory. Describing McCulloch–Pitts nets in this way turned them into a useful tool for designing circuits for computing mechanisms. This is

how von Neumann would later use them (von Neumann 1945). If this were all there was to it, which functions are computable by McCulloch–Pitts nets would be an innocent technical question devoid of epistemological significance.

But one intriguing aspect of the above passage about nets and computation is the way it relates McCulloch–Pitts nets to the Church-Turing thesis (CT). Turing and other logicians had justified CT – the thesis that any effectively calculable function is computable by Turing Machines – by intuitive mathematical considerations. In their passage, McCulloch and Pitts offered a ''psychological'' justification for CT, based on the computational limitations of the human brain. Since what their nets can compute (even with the help of ''a tape, scanners, and suitable efferents'') can be computed by Turing Machines, they implicitly suggested that what computing humans can effectively calculate can be computed by Turing Machines. By stating that the computational limitations of their nets provide a psychological justification of CT, McCulloch and Pitts presupposed that the computational limitations of their nets capture the computational limitations of brains, and that the computational limitations of brains correspond to the ''psychological'' limitations of humans engaged in computation. If so, then defining computable functions in terms of Turing Machines (or any other computationally equivalent formalism) is justified. McCulloch and Pitts seemed to believe that ordinary human computations using pencil and paper (which is what CT is about), and more generally the ''pulling of consequences out of premises,'' could be explained directly in terms of the computations performed by their nets. Thus McCulloch and Pitts attributed epistemological significance to the fact that their nets compute.

Indeed, the main purpose of their theory was to account for mental functions, such as computation and inference, by proposing possible neural-like mechanisms. As McCulloch explained a few years later, he and Pitts were interpreting neural inputs and outputs as if they were symbols written on the tape of a Turing Machine:

> What we thought we were doing (and I thought we succeeded fairly well) was treating the brain as a Turing machine... The important thing was, for us, that we had to take a logic and subscript it for the time of occurrence of a signal (which is, if you will, no more than a proposition on the move). This was needed in order to construct theory enough to be able to state how a nervous system could do anything. The delightful thing is that the very simplest set of appropriate assumptions is sufficient to show that a nervous system can compute any computable number. It is that kind of a device, if you like a Turing machine.[90]

In comparing brains to Turing Machines, McCulloch thought that by constructing their theory, they showed how brains "could do anything," including performing "the kind of [mental] functions which a brain must perform if it is only to go wrong and have a psychosis."[91] "Treating the brain as a Turing machine" was a crucial part of McCulloch and Pitts's attempt at solving the mind–body problem.

8. "CONSEQUENCES"

McCulloch and Pitts ended their paper by drawing what they called "consequences," in a section that introduced several metaphysical and epistemological themes and related them to McCulloch–Pitts nets. This final section of their paper demonstrates the wide scope of their theory, provides further context to the theory, and hence deserves to be followed it in its entirety. It starts with a general point about the causal structure of nets, which is such that, from a given event in a net, it may be impossible to infer either its cause or the time of its cause's occurrence:

> Causality, which requires description of states and a law of necessary connections relating them, has appeared in several forms in several sciences, but never, except in statistics, has it been as irreciprocal as in this theory. Specification for any one time of afferent stimulation and of the activity of all constituent neurons, each an "all-or-none" affair, determines the state. Specification of the nervous net provides the law of necessary connection whereby one can compute from the description of any state that of the succeeding state, but the inclusion of disjunctive relations prevents complete determination of the one before. Moreover, the regenerative activity of constituent circles renders reference indefinite as to time past.[92]

From this relatively straightforward observation about the causal structure of McCulloch–Pitts nets, they drew striking epistemological conclusions:

> Thus our knowledge of the world, including ourselves, is incomplete as to space and indefinite as to time. This ignorance, implicit in all our brains, is the counterpart of the abstraction which renders our knowledge useful. The role of brains in determining the epistemic relations of our theories to our observations and of these to the facts is all too clear, for *it is apparent that every idea and every sensation is realized by activity within that net*, and by no such activity are the actual afferents fully determined.[93]

This passage makes it clear that McCulloch and Pitts thought of individual neuronal pulses and their relations as realizations of sensations, ideas, and their epistemic relations. This assumption – which

had been introduced in the paper by an oblique reference to "considerations impertinent to this argument" – allowed them to draw conclusions about epistemological limitations of the mind directly from the causal structure of their nets (assuming also that brains instantiate the relevant features of nets, namely disjunctive connections and closed loops of activity).

The next passage drew further epistemological conclusions. McCulloch and Pitts noted that changing a network after a stimulus was received would introduce further difficulties in inferring the stimulus from the net's current activity, impoverishing the subject's knowledge and leading to cognitive dysfunctions:

> There is no theory we may hold and no observation we can make that will retain so much as its old defective reference to the facts if the net be altered. Tinnitus, paraesthesias, hallucinations, delusions, confusions and disorientations intervene. Thus empiry [i.e., experience] confirms that if our nets are undefined, our facts are undefined, and to the "real" we can attribute not so much as one quality or "form."[94]

It is worth recalling that McCulloch and Pitts reduced nets that change over time to nets of fixed structure. In their theory, a net's fixed structure determines what a subject can infer about the external world from the net's current activity, so any change in the net's structure diminishes the subject's knowledge and hence is dysfunctional. If the net's structure remains fixed and if its past activity is known, however, it is possible to know precisely which patterns of stimulation gave rise to the net's current activity.[95] Perhaps because of this, McCulloch and Pitts thought they had something to say on the Kantian theme that the mind can know only phenomena not things in themselves: "With determination of the net, the unknowable object of knowledge, the 'thing in itself,' ceases to be unknowable."[96] This statement is conspicuous but unclear. It seems to suggest that if a subject can know the structure and past activity of her own net, then she can know things in themselves.

After drawing their epistemological consequences, McCulloch and Pitts went on to offer some morals to the psychologists. They started with two points: first, they stated a reductionist doctrine according to which their theory had the resources to reduce psychology to neurophysiology; second, they argued that, because of the all-or-none character of neural activity, the most fundamental relations among psychological events are those of two-valued logic. In making their case about psychology, McCulloch and Pitts stated very explicitly

that they interpreted nervous activity as having "intentional character":

> To psychology, however defined, specification of the net would contribute all that could be achieved in that field – even if the analysis were pushed to ultimate psychic units or "psychons," for a psychon can be no less than the activity of a single neuron. Since that activity is inherently propositional, all psychic events have an intentional, or "semiotic," character. The "all-or-none" law of these activities, and the conformity of their relations to those of the logic of propositions, insure that the relations of psychons are those of the two-valued logic of propositions. Thus in psychology, introspective, behavioristic or physiological, the fundamental relations are those of two-valued logic.[97]

The long final paragraph begins with a summary of the "consequences" and a restatement that mental phenomena are now derivable from neurophysiology:

> Hence arise constructional solutions of holistic problems involving the differentiated continuum of sense awareness and the normative, perfective and resolvent properties of perception and execution. From the irreciprocity of causality it follows that even if the net be known, though we may predict future from present activities, we can deduce neither afferent from central, nor central from efferent, nor past from present activities – conclusions which are reinforced by the contradictory testimony of eyewitnesses, by the difficulty of diagnosing differentially the organically diseased, the hysteric and the malingerer, and by comparing one's own memories or recollections with his [*sic*] contemporaneous records. Moreover, systems which so respond to the difference between afferents to a regenerative net and certain activity within that net, as to reduce the difference, exhibit purposive behavior; and organisms are known to possess many such systems, subserving homeostasis, appetition and attention. Thus both the formal and the final aspects of that activity which we are want to call *mental* are rigorously deducible from present neurophysiology.[98]

The same paragraph continues with "consequences" relevant to psychiatry. One is that – contrary to the teachings of psychoanalysis – knowing the history of a patient is unnecessary for treating mental illness. A more general one is that mental diseases reduce to properties of neural nets, and even more generally that the mind–body problem is solved. McCulloch and Pitts were giving a direct answer to Sherrington's statement, mentioned in Section 3, that "mind goes more ghostly than a ghost" (Sherrington 1940):

> The psychiatrist may take comfort from the obvious conclusion concerning causality – that, for prognosis, history is never necessary. He can take little from the equally valid conclusion that his observables are explicable only in terms of nervous activities which, until recently, have been beyond his ken. The crux of this ignorance is that inference from any sample of overt behavior to nervous nets is not unique, whereas, of imaginable nets, only one in fact exists, and may, at any moment, exhibit some

unpredictable activity. Certainly for the psychiatrist it is more to the point that in such systems "Mind" no longer "goes more ghostly than a ghost." Instead, diseased mentality can be understood without loss of scope or rigor, in the scientific terms of neurophysiology.[99]

The essay ends with an appeal to neurology and mathematical biophysics:

For neurology, the theory sharpens the distinction between nets necessary or merely sufficient for given activities, and so clarifies the relations of disturbed structure to disturbed function. In its own domain the difference between equivalent nets and nets equivalent in the narrow sense indicates the appropriate use and importance of temporal studies of nervous activity: and to mathematical biophysics the theory contributes a tool for rigorous symbolic treatment of known nets and an easy method of constructing hypothetical nets of required properties.[100]

The last point, the method of construction of "hypothetical nets of required properties," highlights one the most fruitful legacies of the paper. From then on, McCulloch and Pitts, soon followed by generations of researchers, would use the techniques developed in this paper, and modifications thereof, to design neural networks to explain neural and mental phenomena.

9. CONSEQUENCES

McCulloch and Pitts's project was not to systematize and explain observations about the nervous system – it was to explain knowledge and other mental phenomena in terms of mechanisms that resembled neural ones. To do this, they assumed that mental states can be analyzed in terms of mental atoms endowed with propositional content, the psychons, and that the neural correlates of mental phenomena correspond to precise configurations of neuronal pulses: individual pulses correspond to individual psychons, and causal relations among pulses correspond to inferential relations among psychons.

McCulloch and Pitts's paper offered a mathematical technique for designing neural nets to implement certain inferential relations among propositions, and suggested that those inferences are mathematically equivalent to certain computations. The paper didn't mention computers, because modern computers didn't exist yet. Nonetheless, their technique for nets without circles could be used in designing circuits for digital computers, because it allowed the design of circuits that compute any desired Boolean function. Since circuits

computing Boolean functions became the building blocks of modern digital computers, McCulloch and Pitts's technique got co-opted by von Neumann (1945) as part of what is now called logic design.

In the 1950s, the question raised by McCulloch and Pitts about what their nets (with or without circles) could compute led to the development of finite automata, one of the most important formalisms in the theory of computation.

McCulloch and Pitts's nets were ostensibly "neural" in the sense that the *on* and *off* values of their units were inspired by the all-or-none character of neuronal activity. However, McCulloch–Pitts nets were heavily simplified and idealized relative to the then known properties of neurons and neural nets. The theory did not offer testable predictions or explanations for observable neural phenomena. It was quite removed from what neurophysiologists could do in their labs. This may be why neurophysiologists largely ignored McCulloch and Pitts's theory. Even McCulloch and Pitts, in their later empirical neurophysiological work, did not make direct use of their theory.

But McCulloch and Pitts's theory found a sympathetic audience in people, such as Norbert Wiener and John von Neumann, who were interested in epistemology but trained in mathematics or engineering more than in neurophysiology. For one thing, these scientists liked the claim that the mind had been reduced to the brain; today, some of their intellectual heirs still see the solution to the mind–body problem as McCulloch and Pitts's great contribution.[101] For another thing, they liked the operationalist flavor of the theory, whereby the design of nets was seen as all there was to the performance of inferences and more generally to mental phenomena.[102] Most of all, they liked the mathematical tools and what they saw as their potential for building intelligent machines. They started exploring the technique offered by McCulloch and Pitts. The mathematical techniques got elaborated, modified, and enriched, but the goal remained to explain knowledge in particular and the mind in general using "computing" mechanisms. As Margaret Boden has argued, McCulloch and Pitts's theory was the common origin of both the connectionist and the classical approach to computational artificial intelligence.[103]

McCulloch and Pitts's theory was not the only source of modern computational theories of mind and brain. But McCulloch and Pitts's use of computation to describe neural functions, together with their proposal to explain mental phenomena directly in terms of neural computations, contributed to a large shift in the use of computation

that occurred around the middle of the 20th century. Before 1943, computing was thought of as one human activity among others (e.g., cooking, walking, or talking). After 1943, computing could be thought of as, in a sense, all that humans did. Under McCulloch and Pitts's theory, any net could be described as performing a computation. So in the sense in which McCulloch–Pitts nets compute, and to the extent that McCulloch–Pitts nets are a good model of the brain, every neural activity is a computation. Given that McCulloch and Pitts considered the computations of their nets to be explanations of mental processes and human behavior, every mental process was turned into a computation, and every behavior into the output of a computation. As a further consequence, the Church-Turing thesis (CT) was turned from a thesis about what functions can be effectively calculated into a thesis about the power and limitations of brains.

If subsequent scientists and philosophers had realized that this shift in the notion of computation and in the interpretation of CT was an effect of McCulloch and Pitts's theory, with its idealizations and simplifications of neurons, and with its assumptions about the computational and epistemological significance of neural activity, this would be unproblematic. The problem is that after 1943, many took McCulloch and Pitts's reinterpretation of neural activity and CT without question, and thought it based on mathematically proven facts about brains. Invoking CT in support of computational theories of minds and brains became commonplace. In this changed context, it became natural for many people to read even Turing's own argument for CT (in Turing 1936–37) as a defense of computationalism.[104]

After McCulloch and Pitts's paper, the idea that CT is somehow a psychological thesis about human cognitive faculties, or perhaps a methodological restriction on psychological theories, would stick and would be used time and again to justify computational theories of mind and brain. For example, von Neumann made statements that resembled this interpretation of CT.[105] Another idea would be that since the brain can only do what is computable, there is a computational theory of the brain (e.g. Webb 1980). I find it ironic that McCulloch and Pitts made many of their simplifying assumptions about networks of neurons in order to solve the mind–body problem by using logic and Turing Machines as descriptions of the nervous system, but after their theory was formulated, their theory was used as evidence that the brain is indeed a computing mechanism.

In the years following 1943, many mathematically inclined neuroscientists reverted to modeling neurons and neural nets using

differential and integral equations, which were the tools pioneered by Rashevsky, and supplemented Rashevsky's method with concepts and techniques taken from mathematical physics, control theory, probability and statistics, and information theory. Even so, the legacy of McCulloch and Pitts's computational theory is still at work in current theoretical neuroscience, at least in the terminological choice to describe neural activity as computation. Even though typically, current mathematical models do not employ logic or computability theory to describe or explain neural systems, many theoretical neuroscientists still describe neural nets, neurons, and even sub-neuronal structures such as dendrites and synapses, as performing computations.[106]

I have argued that McCulloch and Pitts's computational theory rested on two principal moves, both of which are problematic. On the one hand, they simplified and idealized neurons so that propositional inferences could be mapped onto neural events and *vice versa*. On the other hand, they assumed that neural pulses correspond to atomic mental events endowed with propositional content. McCulloch and Pitts seemed to suggest that their first move justified the second, which is dubious at best. Furthermore, theoretical neuroscientists later replaced McCulloch and Pitts's theory of neural nets with more empirically adequate models, which were no longer based on a direct description of neural events in terms of propositional inferences. But in spite of the difficulties, both empirical and conceptual, with McCulloch and Pitts's way of ascribing computations to the brain, the computational theory of mind and brain took on a life of its own. McCulloch and Pitts's views – that neural nets perform computations (in the sense of computability theory) and that neural computations explain mental phenomena – stuck and became the mainstream theory of brain and mind. It may be time to rethink the extent to which those views are justified in light of current knowledge of neural mechanisms.

ACKNOWLEDGMENTS

A version of this paper was presented at Washington University in St. Louis, MO; I am grateful to the audience for their feedback. Thanks to Peter Bradley, Carl Craver, Peter Machamer, Sam Scott, and two referees for their helpful comments. Special thanks to Ken Aizawa, whose generous and informed comments led to an extensive revision of the paper and many improvements.

NOTES

[1] For surveys of this literature, see Rashevsky (1938, 1940) and Householder and Landahl (1945). See also Aizawa (1992, 1996) and Abraham (forthcoming).
[2] Turing (1947, 111, 123).
[3] Turing (1948, 5, 1950, 451).
[4] Turing and McCulloch discussed about computing with some of the same people, such as John von Neumann, Norbert Wiener, and Claude Shannon (Hodges 1983; Aspray 1985; Heims 1991). McCulloch and Pitts's 1943 paper was discussed among those people shortly after it was published, so Turing would likely have heard about it. More specifically, Turing's report on an Automatic Computing Engine (Turing 1945) cites von Neumann's "First Draft of a Report on the EDVAC" (von Neumann 1945), and according to Andrew Hodges, in 1946 Turing was using the "notation for logical networks" introduced by von Neumann 1945 (Hodges 1983, 343). As we shall see, von Neumann explicitly acknowledged McCulloch and Pitts's work as the source of his notation. At the very least, Turing would have known about McCulloch and Pitts's work from studying von Neumann's paper.
[5] Koch and Segev (2000, 1171).
[6] Some technical aspects of the paper are discussed in Perkel (1988), Arbib (1989), Cowan (1990a, b, c), and Díaz and Mira (1996).
[7] For more on McCulloch and Pitts's background, and the events that led to the formulation of McCulloch and Pitts's theory, see Frank (1994), Arbib (2000), Smalheiser (2000), Kay (2001), and Abraham (2002, 2003). See also Piccinini (2003a, Chaps. 2 and 3).
[8] McCulloch (1961, 8).
[9] On McCulloch's early psychological theory, see McCulloch (1961, 8–9, 1965, 392–393); and Abraham (2002, 7).
[10] Biographical Sketch of Warren S. McCulloch, ca. 1948. Warren S. McCulloch Papers, Series II, Box 11, file folder Curriculum Vitae. The Warren S. McCulloch Papers are at the Library of the American Philosophical Society, Philadelphia, PA.
[11] McCulloch (1974, 30).
[12] For an account of these events, see McCulloch (1974, 30–31).
[13] McCulloch (1974, 32).
[14] Biographical Sketch of Warren S. McCulloch, ca. 1948. Warren S. McCulloch Papers, Series II, Box 11, file folder Curriculum Vitae. The same Biographical Sketch also says that this was the time when McCulloch "attempted to make sense of the logic of transitive ver[b]s," which conflicts with what he wrote in his later autobiographical essays. Given the lack of primary sources and given McCulloch's inconsistencies in his autobiographical writings, it is hard to date his early work accurately. But in spite of some inconsistencies with dates, in all his relevant writings McCulloch emphasized his early interest in logic and his attempts to apply logic to psychology and later to a theory of the brain. It is thus hard to believe Lettvin when he wrote that until McCulloch worked with Pitts in the early 1940s, McCulloch had not applied "Boolean logic" to the working of the brain (Lettvin 1989a, 12). Since Lettvin met McCulloch only around 1940, Lettvin may never have discovered McCulloch's early efforts in this direction.
[15] For an account of these events, see McCulloch (1961, 1974, 30–32) and Arbib (2000, 213).

[16] McCulloch (1940, 271).
[17] McCulloch's many publications on neurophysiology are reprinted in his *Collected Works* (McCulloch 1989). For more on McCulloch's work in neurophysiology, see Abraham (2003).
[18] Heims (1991, 34ff).
[19] Letter by McCulloch to Ralph Lillie, ca. February 1943. Warren S. McCulloch Papers, Series 1, Box 12, file folder Lillie.
[20] Ibid.
[21] Northrop (1940, 128) cited by Abraham (2002, 6).
[22] For a more detailed account of Northrop's philosophy of science, see Abraham (2002, 6–7).
[23] For more on Turing's contributions, their relevance, and Turing's views on intelligence, see Piccinini (2000, 2003b).
[24] Von Neumann (1951, 33).
[25] Rashevsky (1935, 1936a, b, c, 1938). On Rashevsky and his group, see Abraham (2002, 13–18); Abraham (forthcoming); Aizawa (1996).
[26] Rashevsky (1938, vii).
[27] Ibid., 9.
[28] Ibid., 9.
[29] Ibid., 10.
[30] Rashevsky (1938, Part I).
[31] Ibid., Part II.
[32] Rashevsky (1936a; 1938, Part III).
[33] Rashevsky (1938, 217).
[34] See, respectively, Rashevsky (1938, Chaps. XXV, XXVI, XXVII, XXIX).
[35] See, respectively, Rashevsky (1940, Chaps. X, XI, XII).
[36] Householder (1941a, b, c, 1942).
[37] Householder (1941a, 63).
[38] ibid., 64.
[39] Accounts of Pitts's life also contain fictionalized stories, apparently propagated by McCulloch. Smalheiser gives a nice summary of Pitt's life, work and personality. An important source on Pitts is his "life-long friend" Lettvin (1989a, b).
[40] Heims (1991, 40), Smalheiser (2000, 219). Letter by Lettvin to Wiener, dated ca. April, 1946. Norbert Wiener Papers, Box 4, file folder 70.
[41] Pitts (1942a, 129), Pitts (1943a, 31).
[42] Pitts (1942a, b, 1943a).
[43] Pitts (1942a, 121); italics in the original.
[44] Ibid.
[45] Pitts (1943a, 23).
[46] Pitts (1943b, c).
[47] Pitts (1943b, 2).
[48] Lettvin and Pitts (1943).
[49] Jerome Lettvin reported their ironic intent to MIT archivist Jeff Mifflin (personal communication).
[50] Lettvin interview, in Anderson and Rosenfeld (1998, 3).
[51] Lettvin (1989a, 10).
[52] Lettvin was not always accurate in his historical judgments. For some clear cases, see fns. 14 and 87.

[53] Shortly before both of them died, Pitts wrote McCulloch from his hospital bed, commenting in detail on their conditions and expressing the wish that they meet again and talk about philosophy. Letter by Pitts to McCulloch, dated April 21, 1969. Warren S. McCulloch Papers, Series I, Box 16, file folder Pitts.
[54] McCulloch (1974, 36).
[55] Landahl et al. (1943). See also Householder and Landahl (1945).
[56] Discussion by Dr. McCulloch of a paper by Dr. Alexander on Fundamental Concepts of Psychosomatic Research, Illinois Psychiatric Society, dated May 22, 1943. Warren S. McCulloch Papers, Series II, Box 28, file folder On Dr. Alexander's paper.
[57] Ibid. The works referenced by McCulloch are Sherrington (1940) and Köhler (1938).
[58] Ibid.
[59] The relationship between computational theories of mind and brain and theories of mental content, in McCulloch and other computationalists, is analyzed in detail in Piccinini (forthcoming a).
[60] Letter by McCulloch to Fremont-Smith, dated June 24, 1942. Warren S. McCulloch Papers, Series I, Box 7, file folder Fremont-Smith.
[61] Letter by McCulloch to Ralph Lillie, ca. February 1943. Warren S. McCulloch Papers, Series 1, Box 12, file folder Lillie.
[62] Although McCulloch formulated his theory so as to ascribe content to mental states, he did not give an explicit formulation of the problem of how mental states can have content, and never gave an explicit solution to that problem. The problem of mental content started being explicitly discussed in the literature on computational theories of mind only in the late 1960s. For a reconstruction of how that came about, see Piccinini (forthcoming a).
[63] Fitch (1944).
[64] According to Lettvin, an important source of the logic gate model of the neuron was the recent discovery by David Lloyd of direct excitation and inhibition between single neurons: "it was not until David Lloyd's work in 1939–41 that the direct monosynaptic inhibitory and excitatory actions of nervous pulses were demonstrated. This finding, more than anything else, led Warren and Walter to conceive of single neurons as doing logical operations (a la Leibnitz and Boole) and acting as gates" (Lettvin's 1988, Foreward to the second edition of *Embodiments of Mind*, cited by Heims 1991, 233–234). In light of McCulloch's professions of belief in his logical conception of the nervous system since the early 1930s, it is unlikely that Lloyd's work motivated McCulloch and Pitts's theory other than by providing experimental validation for some of their ideas. As Ken Aizawa has pointed out to me, not only did McCulloch and Pitts not cite Lloyd's work, but Lettvin himself once stated in conversation with Aizawa that Lloyd's work had "nothing" to do with McCulloch and Pitts's (1943) paper.
[65] McCulloch and Pitts (1943, 19–21).
[66] Ibid., 21; emphasis added.
[67] Ibid., 22.
[68] Ibid.
[69] Ibid., 29–30.
[70] Ibid., 20. Neurons also exhibit spontaneous activity, that is, activity in the absence of stimuli.

[71] Ibid., 21. Today, many consider synaptic plasticity the neural mechanism for learning and memory. But although speculative explanations of learning and memory in terms of changes in neuronal connections can be traced back to the late 19th century (Breidbach 2001), synaptic plasticity did not start to be observed in the laboratory until the 1950s (Craver 2003).
[72] Ibid., 29–30.
[73] Ibid., 22.
[74] McCulloch (1974, 36).
[75] Letter by McCulloch to Ralph Lillie, ca. February 1943. Warren S. McCulloch Papers, Series 1, Box 12, file folder Lillie.
Cf. also Lettvin:
The *Logical Calculus*, McCulloch knew, was not even a caricature of any existing nervous process. Indeed he made that very clear at the time of writing. But is [*sic*] was a possible and useful assembly of axiomatized neurons, and that seemed to him a far greater accomplishment than a true description of any definitely known neuronal circuit (of which none then existed) (Lettvin 1989b, 518).
[76] McCulloch and Pitts (1943, 24).
[77] Householder (1941a, 64).
[78] Pitts (1943a, 23).
[79] Their actual proof was not quite a mathematical induction because they didn't show how to combine nets of arbitrary size, but the technical details are unimportant here.
[80] For more on this, and more details about the immediate reception and historical relevance of McCulloch and Pitts's theory, see Piccinini (2003a, Chaps. 5 and 6).
[81] For more on logic design and computer design in general, see Patterson and Hennessy (1998). For a detailed philosophical account of computing mechanisms, see Piccinini (2003a, Chap. 10).
[82] Ibid., 27.
[83] Ibid., 28.
[84] Ibid., 28.
[85] Ibid., 34.
[86] Every commentator points this out, starting with Fitch (1944, 51). See also Arbib (1989). McCulloch and Pitts's mathematical treatment of their nets was superseded a few years later by Kleene's treatment (Kleene 1956; see below).
[87] Ibid., 35. This reference to "a Turing machine" and to "the Turing definition of computability" proves that both McCulloch and Pitts knew of Turing's 1936–37 work. Lettvin was thus mistaken in stating that at the time they wrote their 1943 paper, "neither Warren [McCulloch] nor Walter [Pitts] knew of Turing's paper of 1937" (Lettvin 1989a, 515).
[88] Letter by McCulloch to Schouten, dated October 18, 1948. Warren S. McCulloch Papers, Series I, Box 17, file folder Schouten.
[89] Kleene (1956, 4, 17, 22).
[90] This statement is attributed to McCulloch in a discussion published in von Neumann (1951, 32–33).
[91] Ibid.
[92] McCulloch and Pitts (1943, 35).
[93] Ibid., 35–37, emphasis added.
[94] Ibid., 37.

[95] This is the case when either there are no disjunctive connections in a net, or one knows the activity of the neurons that send inputs to the disjunctive connections in a net.
[96] Ibid., 37.
[97] Ibid., 37–38.
[98] Ibid., 38.
[99] Ibid., 38.
[100] Ibid., 38–39.
[101] Arbib (2000, 212, and 213), Lettvin (1989b, 514). McCulloch and Pitts's theory was also important motivation behind what came to be known in philosophy as the functionalist solution to the mind-body problem and early computational theories of mind that were associated with functionalism. That story is told in Piccinini (forthcoming b).
[102] Cf. Wiener (1948, 147), Shannon and McCarthy (1956).
[103] Boden (1991). For more on the impact of McCulloch and Pitts's work on early computationalism in cybernetics and artificial intelligence, see also Piccinini (2003a, Chaps. 5 and 6).
[104] Turing's argument is interpreted in this way by, e.g., Cleland (1993, 284), Shanker (1995, 55), Fodor (1998), and Webb (1980). The question of the exact relationship between computationalism and CT is addressed at length in Piccinini (2003a, Chapter 7). Cf. also Copeland (2000, 2002) and Piccinini (2003b).
[105] von Neumann (1951). For more recent examples, see Chalmers (1996), Churchland and Churchland (1990), Dennett (1978), Fodor (1981), Haugeland (1981), McGee (1991), and Pylyshyn (1984).
[106] For an introduction to contemporary theoretical neuroscience, see Dayan and Abbott (2001). See also Churchland and Sejnowski (1992), Koch (1999), and Rieke et al. (1997).

REFERENCES

Abraham, T. H.: 2002, '(Physio)logical Circuits: The Intellectual Origins of the McCulloch–Pitts Neural Networks', *Journal of the History of the Behavioral Sciences* **38**(1), 3–25.

Abraham, T. H.: 2003, 'Integrating Mind and Brain: Warren S. McCulloch, Cerebral Localization, and Experimental Epistemology', *Endeavour* **27**(1), 32–38.

Abraham, T. H.: forthcoming, 'Nicolas Rashevsky's Mathematical Biophysics', *Journal of the History of Biology*.

Aizawa, K.: 1996, 'Some Neural Network Theorizing Before McCulloch: Nicolas Rashevsky's Mathematical Biophysics', in R. Moreno Díaz and J. Mira (eds.), *Brain Processes, Theories, and Models: An International Conference in Honor of W. S. McCulloch 25 Years after His Death*, MIT Press, Cambridge, MA, pp. 64–70.

Arbib, M. A.: 1989, 'Comments on ''A Logical Calculus of the Ideas Immanent in Nervous Activity'', in R. McCulloch (ed.), *Collected Works of Warren S. McCulloch*, Intersystems, Salinas, CA, pp. 341–342.

Arbib, M. A.: 2000, 'Warren McCulloch's Search for the Logic of the Nervous System', *Perspectives in Biology and Medicine* **43**(2), 193–216.

Aspray, W.: 1985, 'The Scientific Conceptualization of Information: A Survey', *Annals of the History of Computing* **7**(2), 117–140.

Boden, M.: 1991, 'Horses of a Different Color?' in W. Ramsey, S. P. Stich and D. E. Rumelhart (eds.), *Philosophy and Connectionist Theory*, LEA, Hillsdale, pp. 3–19.

Breidbach, O.: 2001, 'The Origin and Development of the Neurosciences', in P. Machamer, R. Grush and P. McLaughlin (eds.), *Theory and Method in the Neurosciences*, University of Pittsburgh Press, Pittsburgh, PA, pp. 7–29.

Chalmers, D. J.: 1996, *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, Oxford.

Churchland, P. S. and T. J. Sejnowski: 1992. *The Computational Brain*. MIT Press, Cambridge, MA.

Cleland, C. E.: 1993, 'Is the Church–Turing Thesis True?' *Minds and Machines* **3**, 283–312.

Copeland, B. J.: 2000, 'Narrow Versus Wide Mechanism: Including a Re-Examination of Turing's Views on the Mind–Machine Issue', *The Journal of Philosophy* **XCVI**(1), 5–32.

Copeland, B. J.: 2002, 'The Church–Turing Thesis', in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Fall 2002 Edition)*, URL = <http://plato.stanford.edu/archives/fall2002/entries/church-turing/>.

Cowan, J. D.: 1990a, 'McCulloch–Pitts and Related Neural Nets from 1943 to 1989', *Bulletin of Mathematical Biology* **52**(1/2), 73–97.

Cowan, J. D.: 1990b, 'Neural Networks: The Early Days', in D. S. Touretzky (ed.) *Advances in Neural Information Processing Systems 2*, Morgan Kaufmann, San Mateo, CA, pp. 829–842.

Cowan, J. D.: 1990c, 'Von Neumann and Neural Networks', in J. Glimm, J. Impagliazzo and I. Singer (eds.), *The Legacy of John von Neumann*, American Mathematical Society, Providence, pp. 243–274.

Craik, K. J. W.: 1943, *The Nature of Explanation*. Cambridge University Press, Cambridge.

Craver, C.: 2003, 'The Making of a Memory Mechanism', *Journal of the History of Biology* **36**, 153–195.

Dayan, P. and L. F. Abbott: 2001, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*, MIT Press, Cambridge, MA.

Fitch, F.: 1944, 'Review of McCulloch and Pitts 1943', *Journal of Symbolic Logic* **9**(2), 49–50.

Fodor, J. A.: 1981, *Representations*, MIT Press, Cambridge, MA.

Fodor, J. A.: 1998, *Concepts*, Clarendon Press, Oxford.

Frank, R: 1994, 'Instruments, Nerve Action, and the All-or-None Principle', *Osiris* **9**, 208–235.

Heims, S. J.: 1991, '*Constructing a Social Science for Postwar America: The Cybernetics Group, 1946–1953*, MIT Press, Cambridge, MA.

Haugeland, J.: 1981, 'Analog and Analog', *Philosophical Topics* **12**, 213–225.

Hodges, A.: 1983. *Alan Turing: The Enigma*, Simon and Schuster, New York.

Householder, A. S.: 1941a, 'A Theory of Steady-State Activity in Nerve-Fiber Networks: I. Definitions and Preliminary Lemmas', *Bulletin of Mathematical Biophysics* **3**, 63–69.

Householder, A. S.: 1941b, 'A Theory of Steady-State Activity in Nerve-Fiber Networks II: The Simple Circuit', *Bulletin of Mathematical Biophysics* **3**, 105–112.

Householder, A. S.: 1941c, 'A Theory of Steady-State Activity in Nerve-Fiber Networks III: The Simple Circuit in Complete Activity', *Bulletin of Mathematical Biophysics* **3**, 137–140.

Householder, A. S.: 1942, 'A Theory of Steady-State Activity in Nerve-Fiber Networks IV: *N* Circuits with a Common Synapse', *Bulletin of Mathematical Biophysics* **4**, 7–14.

Householder, A. S. and H. D. Landahl: 1945, *Mathematical Biophysics of the Central Nervous System*, Principia, Bloomington.

Kay, L.: 2001, 'From Logical Neurons to Poetic Embodiments of Mind: Warren S. McCulloch's Project in Neuroscience', *Science in Context* **14**(4), 591–614.

Kleene, S. C.: 1956, 'Representation of Events in Nerve Nets and Finite Automata', in C. E. Shannon and J. McCarthy (eds.), *Automata Studies*, Princeton University Press, Princeton, NJ, pp. 3–42.

Koch, C.: 1999. *Biophysics of Computation: Information Processing in Single Neurons*, Oxford University Press, New York.

Koch, C. and I. Segev: 2000, 'The role of single neurons in information processing', *Nature Neuroscience Supplement* **3**, 1171–1177.

Köhler, W.: 1938, *The Place of Value in a World of Fact*, Liveright, New York.

Kubie, L.: 1930, 'A Theoretical Application to some Neurological Problems of the Properties of Excitation Waves which Move in Closed Circuits', *Brain* **53**(2), 166–177.

Landahl, H. D., W. S. McCulloch, and W. H. Pitts: 1943, 'A Statistical Consequence of the Logical Calculus of Nervous Nets', *Bulletin of Mathematical Biophysics* **5**, 135–137.

Lettvin, J. L.: 1989a, 'Introduction', in R. McCulloch (ed.), *Collected Works of Warren S. McCulloch, Vol. 1*. Intersystems, Salinas, CA, 7–20.

Lettvin, J. L.: 1989b, 'Strychnine Neuronography', in R. McCulloch (ed.), *Collected Works of Warren S. McCulloch, Vol. 1*. Intersystems, Salinas, CA, 50–58.

Lettvin, J. L. and W. H. Pitts: 1943, 'A Mathematical Theory of Affective Psychoses', *Bulletin of Mathematical Biophysics* **5**, 139–148.

McCulloch, W. S.: 1940, 'Joannes Gregorius Dusser de Barenne', *Yale Journal of Biology and Medicine* **12**, 743–746.

McCulloch, W. S.: 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. Reprinted in McCulloch 1965, pp. 1–18.

McCulloch, W. S.: 1965, *Embodiments of Mind,* MIT Press, Cambridge, MA.

McCulloch, W. S.: 1974, 'Recollections of the Many Sources of Cybernetics', *ASC Forum* **VI**(2), 5–16.

McCulloch, W. S. and W. H. Pitts: 1943, 'A Logical Calculus of the Ideas Immanent in Nervous Activity', *Bulletin of Mathematical Biophysics* **7**, 115–133. Reprinted in McCulloch 1964, pp. 16–39.

McGee, V.: 1991, 'We Turing Machines Aren't Expected-Utility Maximizers (Even Ideally)', *Philosophical Studies* **64**, 115–123.

Moreno Díaz, R. and J. Mira (eds.): 1996, *Brain Processes, Theories, and Models: An International Conference in Honor of W. S. McCulloch 25 Years after His Death*, MIT Press, Cambridge, MA.

Patterson, D. A. and J. L. Hennessy: 1998, *Computer Organization and Design: The Hardware/Software Interface*, Morgan Kauffman, San Francisco.

Perkel, D. H.: 1988, 'Logical Neurons: The Enigmatic Legacy of Warren McCulloch', *Trends in Neurosciences* **11**(1), 9–12.

Piccinini, G.: 2000, 'Turing's Rules for the Imitation Game', *Minds and Machines* **10**(4), 573–582.

Piccinini, G.: 2003a, *Computations and Computers in the Sciences of Mind and Brain*, Doctoral dissertation, Department of History and Philosophy of Science, University of Pittsburgh, Pittsburgh, PA. URL = <http://etd.library.pitt.edu/ETD/available/etd-08132003-155121/>

Piccinini, G.: 2003b, 'Alan Turing and the Mathematical Objection', *Minds and Machines* **13**(1), 23–48.

Piccinini, G.: (forthcoming a), 'Functionalism, Computationalism, and Mental Contents', *Canadian Journal of Philosophy*.

Piccinini, G.: (forthcoming b), 'Functionalism, Computationalism, and Mental States', *Studies in the History and Philosophy of Science*.

Pitts, W. H.: 1942a, 'Some Observations on the Simple Neuron Circuit', *Bulletin of Mathematical Biophysics* **4**, 121–129.

Pitts, W. H.: 1942b, 'The Linear Theory of Neuron Networks: The Static Problem', *Bulletin of Mathematical Biophysics* **4**, 169–175.

Pitts, W. H.: 1943a, 'The Linear Theory of Neuron Networks: The Dynamic Problem', *Bulletin of Mathematical Biophysics* **5**, 23–31.

Pitts, W. H.: 1943b, 'A General Theory of Learning and Conditioning: Part I', *Psychometrika* **8**(1), 1–18.

Pitts, W. H.: 1943c, 'A General Theory of Learning and Conditioning: Part II', *Psychometrika* **8**(2), 131–140.

Pylyshyn, Z. W.: 1984, *Computation and Cognition*. MIT Press, Cambridge, MA.

Rashevsky, N.: 1935, 'Foundations of Mathematical Biophysics', *Philosophy of Science* **1**, 176–196.

Rashevsky, N.: 1936a, 'Mathematical Biophysics and Psychology', *Psychometrika* **1**(1), 1–26.

Rashevsky, N.: 1936b, 'Physico-Mathematical Methods in Biological and Social Sciences', *Erkenntnis* **6**, 357–365.

Rashevsky, N.: 1936c, 'Physico-mathematical Methods in Biology', *Biological Reviews* **11**, 345–363.

Rashevsky, N.: 1938, *Mathematical Biophysics: Physicomathematical Foundations of Biology*, University of Chicago Press, Chicago.

Rashevsky, N.: 1940, *Advances and Applications of Mathematical Biology*, University of Chicago Press, Chicago.

Rieke, F., *et al*.: 1997. *Spikes: Exploring the Neural Code*, MIT Press, Cambridge, MA.

Rosenblueth, A., N. Wiener, and J. Bigelow: 1943, 'Behavior, Purpose, and Teleology', *Philosophy of Science* **10**, 18–24.

Shanker, S. G.: 1995, 'Turing and the Origins of AI', *Philosophia Mathematica* **3**, 52–85.

Shannon, C. E. and J. McCarthy: 1956, *Automata Studies*. Princeton University Press, Princeton, NJ.

Shapiro, S.: 1998, 'Church's Thesis', in E. Craig (ed.), *Routledge Encyclopedia of Philosophy, V. 2*, Routledge, London, pp. 351–355.

Sherrington, C. S.: 1940, *Man on His Nature*, Cambridge University Press, Cambridge.
Smalheiser, N. R.: 2000, 'Walter Pitts', *Perspectives in Biology and Medicine* **43**(2), 217–226.
Turing, A. M.: 1936–37 [1965], 'On Computable Numbers, with an Application to the Entscheidungsproblem', in M. Davis (ed.), *The Undecidable*, Raven, Ewlett, pp. 116–154.
Turing, A. M.: 1947, 'Lecture to the London Mathematical Society on 20 February 1947', in D. Ince (ed.), *Mechanical Intelligence*. North-Holland, Amsterdam, pp. 87–105.
Turing, A. M.: 1948, 'Intelligent Machinery', in D. Ince (ed.) *Mechanical Intelligence*. North-Holland, Amsterdam, pp. 87–106.
Turing, A. M.: 1950, 'Computing Machinery and Intelligence', *Mind* **59**, 433–460.
von Neumann, J.: 1945, '*First Draft of a Report on the EDVAC*', Technical Report, Moore School of Electrical Engineering, University of Pennsylvania, Philadelphia, PA.
von Neumann, J.: 1951, 'The General and Logical Theory of Automata', in L. A. Jeffress (ed.), *Cerebral Mechanisms in Behavior*. Wiley, New York, pp. 1–41.
Webb, J. C.: 1980, *Mechanism, Mentalism, and Metamathematics*, Reidel, Dordrecht.
Wiener, N.: 1948, *Cybernetics or Control and Communication in the Animal and the Machine*, MIT Press, Cambridge, MA.
Woodger, J. H.: 1937, *The Axiomatic Method in Biology*, Cambridge University Press, Cambridge.

Department of Philosophy
Washington University
Campus Box 1073
One Brookings Dr.
St. Louis, MO 63130-4899
U.S.A.
E-mail: gpiccini@artsci.wustl.edu