Warren S. McCulloch [*]

Why the Mind Is in the Head

As the industrial revolution concludes in bigger and better bombs, an intellectual revolution opens with bigger and better robots. The former revolution replaced muscles by engines and was limited by the law of the conservation of energy, or of mass-energy. The new revolution threatens us, the thinkers, with technological unemployment, for it will replace brains with machines limited by the law that entropy never decreases. These machines, whose evolution competition will compel us to foster, raise the appropriate practical question: "Why is the mind in the head?"

Coming as I do between psyche anatomized and psyche synthesized, I must so define my terms that I can bridge the traditional gulf between mind and body and the technical gap between things begotten and things made.

By the term "mind," I mean ideas and purposes. By the term "body," I mean stuff and process. Stuff and process are familiar to every physicist as mass and energy in space and time, but ideas and purposes he keeps only in the realm of discourse and will not postulate them of the phenomena he observes. In this I agree with him. But what he observes is some sort of order or invariance in the flux of events. Every object he detects in the world is some sort of regularity. The existence of these objects is the first law of science. To detect regularities in the relations of objects and so construct theoretical physics requires the disciplines of logic and mathematics. In these fundamentally tautological endeavors we invent surprising regularities, complicated transformations which conserve whatever truth may lie in the propositions they transform. This is invariance, many steps removed from simple sensation but not essentially different. It is these regularities, or invariants, which I call ideas, whether they are theorems of great abstraction or qualities simply sensed. The reason for excluding them from physics is that they must not be supposed to be either stuff or process in the causal sequences of any part of the world. They are neither material nor efficient. So, to my mind Newton, Planck, and Jeans sin by introducing God as a sort of mind at large in the world to account for physical effects, like the action of gravity at a distance.

But let us now compel our physicist to account for himself as a part of the physical world. In all fairness, he must stick to his own rules and show in terms of mass, energy, space, and time how it comes about that he creates theoretical physics. He must then become a neurophysiologist (that is what happened to me), but in so doing he will be compelled to answer whether theoretical physics is something which he can discuss in terms of neurophysiology (and *that* is what happened to me). To answer "no" is to remain a physicist undefiled. To answer "yes" is to become a metaphysician – or so I am told.

But is that just? The physicist believes entropy to be somehow in or of physical systems. It may or must increase with time. But it is neither material nor efficient, in fact it is a number, namely, the logarithm of the probability of the state. It is, therefore, a measure of the disorder of an ensemble – or collection of systems. Now Norbert Wiener has proposed that information is orderliness and suggests that we measure it by negative entropy, the logarithm of the reciprocal of the probability of the state. Let us, for this

[*] Warren S. McCulloch, in: "Warren S. McCulloch: Embodiments of Mind", The MIT Press, Cambridge, Mass. 1965, pp. 72-87. (first published in: Cerebral Mechanisms in Behavior, L.A. Jeffress (ed.), The Hixon Symposium, Jon Wiley Publ., New York 1951, pp. 42-57)

argument, accept his suggestion. Ideas are then to be construed as information. Sensation becomes entropic coupling between us and the physical world, and our interchange of ideas, entropic coupling among ourselves. Our knowledge of the world, our conversation – yes, even our inventive thought – are then limited by the law that information may not increase on going through brains, or computing machines.

The attempt to quantify the information leads to a search for an appropriate unit, which, in turn, forces us to distinguish between two types of devices. In so-called logical, or digital, contrivances, a number to be represented is replaced by a number of things – as we may tally grain in a barn by dropping a pebble in a jug for each sheaf. The abacus is such a device. The nervous system is par excellence a logical machine. In so-called analogical contrivances a quantity of something, say a voltage or a distance, is replaced by a number of whatnots or, conversely, the quantity replaces the number. Sense organs and effectors are analogical. For example, the eye and the ear report the continuous variable of intensity by discrete impulses the logarithm of whose frequency approximates the intensity. But this process is carried to extremes in our appreciation of the world in pairs of opposites. As Alcmaeon, the first of experimental neurophysiologists, so well observed, "the majority of things human are two" – white-black, sweet-bitter, good-bad, great-small. Our sense organs, detecting regularities the same in all respects save one, create dichotomies and decide between opposites. These the "brain somehow fits together." From this sprang associationalism, culminating in Mill's evolutionary hypothesis that things are similar for us which have occurred together in the experience of our progenitors, and Kapper's law that nervous structures associated in action become associated in position. Neither proposes any mechanism other than random variation and the survival of those in whom the happy concatenation occurred. We inherit a nervous system so structured that we do perceive similarities (or have ideas) and these, not isolated, but conjoined within the system in many useful ways. That synthetic a priori is the theme of all our physiological psychology, learning excepted.

How, in all these processes, can we "quantify" the amount of information? In analogical devices it is best done by examining the numerical component. They all suffer from a peculiar limitation of accuracy which can usually not be pushed beyond one part in a thousand and almost never beyond one part in a million, even in such a simple matter as weighing. Moreover, analogical devices can not be combined in any way to push the decimal point. By measuring carefully the diameter and circumference of a circle, we might analogically estimate the ratio π to six significant figures. With a digital device, say an abacus, we can compute it to any number of places.

What characterizes a digital, or logical, device is that its possible states are separated sharply. In the simplest case there are only two. Wiener proposes that, for these bivalent systems, we define the unit of information as the decision which state it shall occupy. Notice, now, that this system has one degree of freedom – say, go or no-go – until it receives one unit of information, whereupon it has none. Next consider an ensemble of two such systems. It can be in any one of four states, and two units of information are required to match its degrees of freedom. If it were composed of three systems it might be in any one of eight states, and would require three units of information to fix it. Thus the number of possible states is 2 raised to the power which is the number of systems and each unit of information subtracts one from that exponent. Here Wiener's unit of information is exactly the logarithm to the base 2 of the reciprocal of the probability of the state, which, of course, is the negative entropy of an ensemble of bivalent systems. Neurons are bivalent systems.

Let me define "corruption" as the ratio of information in the input to that in the output. Each eye has something like a hundred million photoreceptors, each of which in a given millisecond can emit one or no impulse. In other words, it is an ensemble which can be in any one of $2^{100,000,000}$ possible states, or the amount of information it has is a hundred million units per millisecond.

Now Pitts and I have computed the information in the output of a piano player surpassing any ever known. We have given him a keyboard of a hundred keys, let him strike independently with each finger with any one of ten strengths ten times per second, and let each hand span ten keys. That sounds like a lot of information, but on computing it we find it is only about two units per millisecond.

Recent telephonic devices have sampled waves every thousandth of a second and passed on one pip if the wave was then of a given deviation from the mean, otherwise no pip. These are relayed to a smearing device and heard. It is better than 90 per cent as intelligible as the original voice. Three such pips per millisecond determined by eight possible values of the wave reproduce an orchestra. So much information at most may we hope to convey.

Whether we figure the ratio of input to output from our impossible player or from human speech, the corruption is of the order of a hundred million to one. Part of this corruption is referable to the coupling of our nervous system to our muscles and is avoided in some of the Crustacea. They use axons of several sizes, and by varying the frequency of discharge obtain more degrees of contraction than there are possible synchronous states of the nerves to their muscles. The viscosity of muscle smears the result in time, so that the rate at which impulses can come over the nerve is wasted by the inability of muscle to follow. In us a nerve of a thousand axons can be in 2^{1000} possible states, whereas the muscle, because it can only add tensions, has only a thousand possible states. 1000 is about 2^{10}; so the corruption in passing from nerve to brawn is 100 to 1.

What becomes of all the rest of the information? To answer that, conceive neurons as telegraphic relays. Each one may be tripped by some combination of signals provided these are very nearly synchronous. It detects the coincidence and only then emits a signal to subsequent relays. Now the threshold of the photoreceptors of the eye is always varying. At any one millisecond it may be tripped by a single photon, and, at another, fail to fire in response to many. By connecting many of these to a coincidence detector set to require a reasonable number of impulses simultaneously, we have a signal which corresponds to a statistically significant fraction of its receptors and so we wash out the random variation of threshold. Thus using the relayed information that fails to agree with other information, we achieve a high probability that what goes on through the nervous system does correspond to something in the world. Perhaps it will be clearer to say it this way. The logical probability that a neuron will have an impulse in one millisecond is 1/2, that two neurons of an ensemble in the same millisecond $1/2 \times 1/2$. The chance that both will fire by chance simultaneously is the product of their probabilities separately; that is, it is smaller; 1/4. Therefore, in the nervous system, by repeatedly demanding coincidence we vastly increase the probability that what is in the output corresponds to something in the input. We pay for certainty with information. The eye relays to the brain about the hundredth part of the information it receives. The chance that what it does relay is due to chance is fantastically small, 2^{-100}, a billionth of a billionth of a billionth of a tenth of one per cent.

Here, then, is the first technically important difference between us and robots. In them we cannot afford to carry out any computations, no matter how simple, in a hundred parallel paths and demand coincidence. Consequently, no computing machine is as likely to go right under conditions as various as those we undergo.

Accordingly to increase certainty every hypothesis should be of minimum logical, or a priori, probability so that, if it be confirmed in experiment, then it shall be because the world is so constructed. Unfortunately for those who quest absolute certainty, a hypothesis of zero logical probability is a contradiction and hence can never be confirmed. Its neurological equivalent would be a neuron that required infinite coincidence to trip it. This, in a finite world, is the same as though it had no afferents. It never fires.

In all of this I take it for granted that you are familiar with the all-or-none law of the nervous impulse, with the brevity of latent addition, with the duration of synaptic delay, with the evidence for spatial summation and for inhibition at a synapse, and with the local origin of energy from metabolism, all of which together insure that the principal circuit actions of the nervous system are those of the digital, or logical, kind. My reason for letting time flow in lapses of a millisecond is based on the work of Lorente de Nó, who has given us some of the best measures of the other properties of conduction. It is probable that no neuron can sustain more impulses than a thousand per second, even under the healthiest conditions, and one millisecond will include synaptic delay, or absolute refractory period, or the front of the pulse itself. These permit us to treat the nervous impulse as an atomic event.

But a nervous impulse is also a signal. It is true if what it proposes is true, otherwise it is false. It is false if it arises from any cause other than the adequate, or proper, excitation of the cell. The threshold of the dark-adapted eye for light is about a photon in several seconds. Pressure applied to the eye will evoke impulses, but the energy required is many million times more. Press on the eye and you see light when there is no light. The signals are false. Thus nervous impulses are atomic signals, or atomic propositions on the move. To them the calculus of propositions applies provided each is subscripted for the time of its occurrence and implication given a domain only in the past. In terms of such a calculus applied to nervous nets, Pitts and I have been able to prove that even nets devoid of circles can realize any proposition which is a logical consequence of its input. As this is the most that any net can do it is obviously an adequate theory. We know, of course, that facilitation and extinction occur, and we showed that whatever these can effect can be done digitally, or discretely, by go, no-go devices. In our first essay, we were unable to obtain much more than the calculus of atomic propositions; but, by introducing circles in which a train of impulses patterned after some fact could circulate, we did get existential operators for time past.

This is the argument: In a net in which there are no re-entrant paths a signal anywhere in the net implies a signal in a neuron nearer to receptors, and so backward in time until we arrive at the receptors. The signal here and now in this net implies the signal sent there just then. But once set going, a disturbance in a closed chain implies that there was a signal in its input at some time but does not indicate at what time. In short, the reverberating activity patterned after something that happened retains the form of the happening but loses track of when it happened. Thus it shows that there was some time at which such and such occurred. The "such and such" is the idea wrenched out of time.

It is an eternal idea in a transitory memory wherein the form exists only so long as the reverberation endures. When that ceases, the form is no longer anywhere. Only this kind of memory remains to aged brains in which no new abiding traces can be made and old ones fade. While we are young, use leaves some sort of change, as freshets cut their channels in the hills so that aftercoming waters follow and enlarge their beds. Yet all other forms of memory, including written records, do nothing which cannot also be achieved by mere reverberation, and hence add nothing to the theory.

There are other closed paths important in the origin of ideas, circuits which have "negative feedback." In terms of them reflexes were first defined as actions starting in some part of the body, setting up impulses to the central nervous system, whence they were reflected to those structures in which they arose, and there stopped or reversed the process that gave rise to them. All inverse feedbacks have this in common, that each establishes some particular state of the system, for they bring it back toward that state by an amount which increases with their deviation from that state. They are, as we say, error-operated. The state toward which they return the system is the goal, or aim, or end *in and of* the operation. This is what is meant by function. On these circuits Cannon founded his theory of homeostasis, and Rosenblueth and Wiener their theory of teleological mechanisms.

Any such circuit becomes a servomechanism as soon as the particular state it is to seek can be determined for it. Thus the stretch reflex tends to keep our muscles at constant length, but *that* length is determined for these circuits by more complicated arcs which traverse almost all parts of the central nervous system and require the reflex to seek those states which permit us to stand and move.

One reflex turns the eyes toward anything that enters the visual field. Its path runs from the eye by fibers that bypass the geniculate to enter the superior colliculi upon which they map the visual field. Here local circuits compute the vector from the center of gaze to the center of gravity of the apparition and send this information to the oculomotor nuclei which, in turn, relay orders to the appropriate muscles and turn the eyes so as to decrease that vector. As it reaches zero the eyes come to rest with the apparition centered. This reflex, I am told, will operate even in a man who has lost one-half his visual cortex, if he is dark adapted and a light, unseen by him, is placed in his blind field. If two are placed there the eyes turn toward a position intermediate. Under these conditions but with cortex *intact*, the eyes turn similarly but then snap from spot to spot, for the reflex is then subservient to impulses from the cerebrum. By turning the eyes so as to center the form, the reflex rids the apparition of the gratuitous particularity of the place at which it appeared. Every reflex, by running through a series of intermediate states to that established by it, rids some item to be observed of some fortuitous specificity. In the case of the collicular reflex, it has selected the centered form from among all possible exemplifications. Once in this, the canonical position, the system is ready for the computation of the form. There is little doubt that in us this computation occurs in the cerebral cortex, notably the visual areas. What , happens then as the eyes turn rapidly is a series of on and off signals from most portions of the eye. These serve to clean the slate for the centered form unencumbered by blurring due to motion. The latter we suffer only when the eyes turn slowly; for instance, when we are very tired.

There are negative feedbacks within the brain. One of these resembles the automatic volume control in the radio. It tends to keep constant the sensory input to the cerebral cortex. In so doing it gives us another existential operator, for it detects that there was some intensity such that it was of this-or-that figure. In the case of vision, this circuit follows two other devices serving the same end; namely, the slow adaptation of the retina and the rapid change of pupillary diameter. All together these enable us to detect the form though the intensity of its illumination range through 39 decibels, that is, from faint starlight to full daylight – only we may not look at the sun without closing our eyes.

There are also appetitive circuits with a part of their path, from receptors to effectors, inside us, and the rest through the world outside. They are said to be inverse feedbacks over the target. Given any two inverse feedbacks which working together would destroy us, like swallowing and inhaling, there is built into us some connection between their paths whereby, when both are set going, one stops the other. In the case of learned modes of appetitive behavior, similar inhibitory links must be acquired or we perish. If we have three incompatible circuits in which A dominates B and B dominates C, the chances are equal whether A will dominate C or C, A. We speak of the end-in-operation of the dominant as of greater value. We have even tried to construct scales of value for diverse ends, but, since dominance is sometimes circular, values are not magnitudes of a single kind, and the terms "greater" and "less" are simply inapplicable. What we have called the value anomaly and regarded as evidence of a lack of order or system bespeaks, in fact, order of a kind we had not imagined, and a system tighter knit. Here endeth the psychological blind alley and Plato's theory of the Good. We cannot make one scale of value that predicts choice. Only knowledge of mechanism permits such prophecy.

I drew a circuit to move a figure, given anywhere in a mosaic of relays, to all positions in one direction. From each relay impulses ascended diagonally in the required direction through sheets of relays resembling the original mosaic and so spaced that their constituent relays formed columns perpendicular to the planes. And I set the threshold of all relays so that none would fire except when a slanting impulse coincided with one in the plane of that relay. I brought the output of every relay vertically all the way down to the original mosaic. Now when there is a simultaneous volley of the required figure at the given place in the original mosaic and at the same time a simultaneous volley to all the relays of any one of the sheets above, the figure is reproduced on that sheet by a volley in the relays where the slanting volley hits the sheet. Thence it projects straight down on the original mosaic. This reproduces the figure at a distance which steps off in the direction of the slant by a number of relays proportional to the height of the excited sheet. Now let the figure endure by a series of volleys at its origin, and excite the sheets successively upward, and the figure will be translated step by step from the origin to all possible positions in the direction of the slant. Whatever shape is present in the input to this circuit is preserved in these successive representations and, as the output descends vertically, the shape is translated without distortion. Von Bonin, who had worked with me on the auditory cortex, when he saw the diagram mistook it for a drawing of that cortex. Certainly we had but to replace the relays by pictures of neurons and the similarity was startling. The parallel functions are even more alike. We can center a form seen by turning our eyes, but there is no way we can tune our ears so as to translate a chord up and down the scale. Our brain receives it at a fixed key or pitch. In the primate these pitches map longitudinally on the input to Heschl's gyrus, so spaced that octaves span nearly equal distances. If this, the primary auditory cortex, worked like my circuit it would move the output up and down the axis of pitch while it preserved the interval, and so the chord. Here is an existential operator for chord regardless of pitch. The output asserts that there were pitches such that there was this or that chord.

Is there anything in physiology corresponding to the sequential excitation of the sheets? And, if so, how fast can it complete a cycle? There is the familiar alpha rhythm of the cortex, a shift of voltage that rises and falls through the cortex ten times per second. Although the correspondence may be entirely fortuitous, this is about the rate at which chords can be distinguished – ten per second. Now we need excellent histological studies by the Golgi method to know whether the detailed connections of cells in this area are what the hypothesis requires. These must be made in specified planes to match our physiological data. Because incoming signals and outgoing signals, like the pulse of scansion, ascend and descend through the cortex, when the cortex is at work the sweep of scansion should disappear, as it does, in the twinkle of details.

There are at least two ways that the output of this primary cortex may convey a chord regardless of pitch. There may be an inverse feedback which stops the figure of excitation when it reaches a canonical position along the axis of pitch, but there is no evidence that this exists. The other way is suggested by anatomy. Beneath the receptive layers of the cortex are columns of cells where properly timed impulses may be accumulated through a time equal to the sum of their synaptic delays to coincide upon efferent cells whose axons go to the adjacent, or secondary, auditory cortex.

If they terminate there at random, and if the cells there merely require coincidence to fire, we will have for every chord regardless of pitch a corresponding spot of maximal coincidence. The activity of this spot proposes the required universal, or idea. If we were to excite this spot electrically in waking man at operation, he should report hearing the chord. He does, but unfortunately no one has asked him whether he hears it at some particular pitch. The experiment is difficult because the primary auditory cortex is buried deep, and the secondary adjacent to it almost as deep, in the fissure of Sylvius. Moreover, I have not been able to map well the projection of the primary upon the secondary; and, finally, the interpretation is complicated by a direct projection upon the second with the sequence of pitches reversed. Fortunately these difficulties are not present in the visual area of man or monkey.

I drew a circuit to extract shape regardless of size; and this was mistaken by both von Bonin and Percival Bailey for a schematic representation of the outer strip of Baillarger which makes the visual cortex the "area striata." We start again with a mosaic. Select a point to represent the center of gaze and map the visual field as a set of concentric circles whose radii are proportional to the logarithm of the angles at the eye. From the mosaic let impulses proceed along branching channels spraying outward as they ascend through sheets of relays in which the density of relays decreases but their threshold increases as we go from below upward. From all of these relays let signals rise to corresponding upper layers of relays where coincidence with sweeping pulses is required, and let the signals of these layers converge on relays of low threshold, thence descend to leave the area striata. Now, with the pulse of scansion we shall have successively in this output all possible dilatations and constrictions of any figure in the input. The possibilities are limited by the grain and gross dimensions of the cortex, but these limit input and output equally. Since we have, in the output, all sizes made from a given one it makes no difference which size was given in the input.

Had we not conformed to present knowledge by mapping radial angles by their logarithms we would be compelled to require that the branching ascent of the input take a radial direction, but as it is it may branch nearly equally in all directions. Hence we cannot hope to detect much difference in histological study, even by silver stains of fibers, between sections cut radially and others, tangentially, in the visual cortex, or even in ones parallel to the surface. Thanks to Ramon y Cajal and Lorente de Nó, we know that the anatomical connections are at least sufficient for the theory. Here, as in audition, if the alpha rhythm evinces the scansion, we should be able to see ten forms per second. We can. Faster, they blur, merge, or glide into one another. Moreover, a rise of metabolic rate with fever or hyperthyroidism causes a rise in alpha frequency and a rise in the number of distinguishable frames per second.

When strychnine is applied locally to a spot on the cortex it causes the cells there to fire almost in unison. The fibers leaving that area then carry nearly synchronous volleys of impulses. These can be traced to the ends of the axons if there are enough of them near together. When these axons turn up again into the cortex anywhere, we can detect them there as a sharp change in voltage, the so-called "strychnine spike." When Dusser de Barenne and I strychninized a pinhead spot on the area striata, strychnine spikes appeared at many points in the secondary visual area as if the output from each spot in the area striata were scattered at random in the secondary visual area. Hence, from any particular set of spots in the primary area there will arise by chance some spot of maximum excitation in the secondary area. Activity at this spot implies activity in some figure of spots output by the primary; hence some shape regardless of size. Electrical stimulation of a spot on the primary visual cortex in waking man is reported by him as a blurred circle of light, whereas similar excitation of a spot in the secondary area is reported as a form. Moreover, this form, while it has a position in space, in the sense that he can point at it, has none in the visual field. Nor does it seem to have size there, any more than the recalled image of the moon seems to subtend one particular angle at the eye.

The mechanism we propose for abstracting chord and form is really computing a kind of average, and that average will not be seriously affected by small perturbations of excitation, of threshold, or even of particular connections as long as they are to cells in the right neighborhood. This conforms to clinical findings. A man may have several holes in his visual cortex, as big as or bigger than, the hole in his retina called the optic disk, and, except in a small number of cases, the forms seen will be unaffected. Although in such a case we can map these blind spots, he will not see them and the things seen will appear to be continuous through the blind spots. Scrutiny of the hypothesis even suggests that this process may account for much relational determination, for the four corners of a square in the input would be completed as a square, whereas parts of the sides might well flop, seeming now a maltese cross and now a square, etc. These flops would be the outcome of rivalry between two maxima for dominance over subsequent areas. Thus, for vision, our hypothesis fits well all known facts.

Older schools of physiological psychology and of neurology, guided by atomistic associational doctrines, tended to think in terms of neurons, each of which had one duty, for example, to know squares. This seems to be at least partially true of spots in the secondary visual cortex. Gestalt psychologists have treated the mosaic of relays of the cortex as if it were a field on which sensations mapped synchronously. This seems more likely true of receptors like the retina, for even its cortical replica is bisected by a line down the middle of the field and the halves mapped far asunder. Now it is easy to show that both of these "caricature" the nervous process. We need only note that a nervous net can take any figure in space, requiring an ensemble of a given number of neurons simultaneously, and convert it into a figure of impulses over a single neuron requiring as many relay times as there were neurons in the ensemble, and vice versa. From this alone it is clear that we cannot tell what kind of thing we must look for in a brain when it has an idea, except that it must be invariant under all those conditions in which that brain is having that idea. So far we have considered particular hypotheses of cortical function. They are almost certainly wrong at some point. Because they have already had to fit many disparate data, they are of little a priori probability. They prophesy the outcome of an infinite number of experiments, some of which are almost certain to refute them.

But with respect to the underlying theory, which is merely glorified tautology, there is no such possibility. It is, in fact, little more than a simple application of the theory of groups of transformations. For any figure in the input of a computing machine it is always possible to calculate an output invariant under a group of transformations. We calculate a set of averages, for all members of the group, of numerical values assigned by an arbitrary functional to each transform of the information conceived as the distribution of excitation at all points and times in an appropriate manifold. To define the figure completely under these transformations, we would need a whole manifold of such averages for various functionals, and this manifold would have to have as many dimensions as our original one; but, for practical purposes, we usually need only a few averages. Since in the finite net of relays the number of transformations in finite time is finite, we may use simple sums instead of averages.

This general theory describes all processes of securing invariants, or having ideas, which we have discovered or invented to date; and one mechanism differs from another in the nature of its arbitrary functional. For example, in the cerebral circuits proposed the functional may always assign the value one to any vector in the manifold if the particular point had a signal in the previous relay time, and, if not, assign it zero, whereas, in the reflex circuit for centering an apparition the functional clearly depends upon the figure of excitation in the manifold, and changes as the form centers; in effect, it assigns the value zero to all save the last transform on the cortex.

We may, of course, make the output of any calculator of invariants (or of several of them) the input to another and so have an idea of ideas, which is what Spinoza calls consciousness, and thus get far away from sensation. But our most remote abstractions are all ultimately reducible to primitive atomic propositions and the calculus of the lowest level. The domain of their implication lies only in time past. If their domain extended into the future, our sensations would imply our thoughts and our thoughts imply deeds. They do not, for even if the threshold of every cell in the nervous system were fixed, between the time we conceive an act and the time the impulses reach the motor horn cell, other signals from the world may get there first, and so often thwart us. We note the failure in the fact and are forced to distinguish between what we will and what we shall do. Hence the notion of the will.

But we do guess at things to come. When we run to catch a baseball we run not toward it but toward the place where it will be when we get there to grab it. This requires prediction. We behave as if there were some law compelling the world to act hereafter as it did of yore. Only one of our predictive circuits has been carefully studied by physiologists. It is responsible for optokinetic nystagmus. It has a tendency to persist, which may be seen when a train stops, for it then attributes motion in the opposite direction to the ties and rails. The earmark of every predictive circuit is that if it has operated long uniformly it will persist in activity, or overshoot; otherwise it could not project regularities from the known past upon the unknown future. This is what, as a scientist, I dread most, for as our memories become stored, we become creatures of our yesterdays – mere hasbeens in a changing world. This leaves no room for learning.

Neurons are cheap and plentiful. If it cost a million dollars to beget a man, one neuron would not cost a mill. They operate with comparatively little energy. The heat generated raises the blood in passage about half a degree, and the flow is half a liter per minute, only a quarter of a kilogram calorie per minute for 10^{10}, that is, 10 billion neurons. Von Neumann would be happy to have their like for the same cost in his robots. His vacuum tubes can work a thousand times as fast as neurons, so he could match a human brain with 10 million tubes; but it would take Niagara Falls to supply the current and the Niagara River to carry away the heat. So he is limited to about the thousandth part of man's computer. He has to be very careful to specify in detail which relays are to be connected to a given relay to trip it. That is not the case in human brains. Wiener has calculated that the maximum amount of information our chromosomes can convey would fill one volume of the *Encyclopaedia Britannica*, which could specify all the connections of ten thousand neurons if that was all it had to do. As we have 10^{10} neurons, we can inherit only the general scheme of the structure of our brains. The rest must be left to chance. Chance includes experience which engenders learning. Ramon y Cajal suggested that learning was the growing of new connections.

I do not doubt that the cerebral cortex may be the most important place in primates. But it is certainly the most difficult place to look for change with use. Think of it as a laminated felt of fibers which serve to associate neighboring rough columns of cells nearly a hundred high and linked together vertically by their axons. These columns are then connected to distant columns by axons which dip into the white malter and emerge elsewhere into the cortex. These last connections I have studied for many years but have at best a general picture of how areas are related, certainly nothing that could give the detail necessary to distinguish between its connections before and after learning.

To understand its proper function we need to know what it computes. Its output is some function of its input. As yet we do not know, even for the simplest structure, what that function is. We have only a few input-output curves for the monosynaptic reflex arc obtained by David Lloyd, and now a few more by Arturo Rosenblueth. Walter Pitts is analyzing them mathematically at the present moment and has as yet no very simple answer. There is no chance that we can do even this for the entire cortex. That is why we need such a hypothesis as we have proposed for particular areas, for these may be disproved by records of electrical activity recorded concurrently at a few specified places.

Contrast our ignorance of its proper function with the detailed present knowledge of the projection of the sensory system upon it. For on at least two-thirds of its surface we can map the surface of the body, outside, and, to some extent, inside, so as to assign to every square millimeter of cortex the origin of its specific afferents and through them the exact position of the organs of sense. Beginning last summer, and continuing right now, the surface of the cerebellum, upon which the body maps similarly, is being stimulated and its projection to the so-called motor and sensory cortex, primary and secondary, explored and plotted millimeter by millimeter. Also now the projections of so-called non-specific afferents are receiving similar attention. Thus within a year or so we will know the geometry of its input and will be ready to seek in loci well defined the temporal pattern of its input.

I wish we could say half as much for our knowledge of its output. Since the days of Bubnoff and Heidenhain it has been electrically stimulated and the resultant change in muscle and gland carefully observed and elaborately recorded. But these responses depend upon the state of all subservient circuits which have yet to be analyzed. Hardly a month passes but what we are confronted by surprises. Frequency as well as shape of electrical pulses have been shown to determine the very path of the descending pulses from one and the same cortical focus: For example, volleys of impulses from the so-called face motor cortex, if more than ten or twelve per second, play principally through the nucleus of the seventh nerve upon the muscles of the face; whereas, if less than ten per second they axe relayed almost exclusively through the nucleus of the twelfth to the tongue. Finally, the response to one and the same form of stimulation of a single focus in the motor cortex for one limb is determined both in amplitude and in direction by the motion and by the position of the limb at the time of stimulation. Clearly, to understand the significance of the output of the cerebral cortex we must know, for every subservient structure, the input-output curves. Even that will not be enough, for when several of them form a re-entrant circuit we must know their relations. Until we do so we will be in danger of attributing to the cerebral cortex functions proper to lower structures.

Last, but not least, the cortex is itself part of many re-entrant systems, and what our hypothesis attributes to cortex alone in securing invariants, or having ideas, may well depend upon loops joining it and the thalamus. From all these uncertainties I would turn to something simple as the monosynaptic arc of the stretch reflex and, by procedures far from normal, try to teach it something. It will be difficult, for in it the connections are as certainly determined as in a man-made computing machine; and we will have to break old connections before we can form new ones.

This brings us back to what I believe is the answer to the question: Why is the mind in the head? Because there, and only there, are hosts of possible connections to be formed as time and circumstance demand. Each new connection serves to set the stage for others yet to come and better fitted to adapt us to the world, for through the cortex pass the greatest inverse feedbacks whose function is the purposive life of the human intellect. The joy of creating ideals, new and eternal, in and of a world, old and temporal, robots have it not. For this my Mother bore me.

REFERENCES

1. Barker, S. H., and Gellhom, E. Influence of suppressor areas on afferent impulses, J. Neurophysiol., 1947, 10, 125-132.

2. Bell, Charles, On the nervous circle which connects the voluntary muscles with the brain, Proc. Roy. Soc 1826, 2, 266-267.

3. Cajal, Ramon y, S. Histologie du système nerveux. 2 vols. Paris: Maloine, 1909, 1911.

4. Lorente de Nó, R. Sections, in: J. F. Fulton's Physiology of the nervous system. London: Oxford University Press, 1943.

5. Maxwell, C., On governors, Proc. Roy. Soc., 1868, 16, 270-283.

6. McColl, H., Servo-mechanisms, New York: D. Van Nostrand Co., 1945.

7. McCulloch, W. S., A heterarchy of values determined by the topology of nerve nets, Bull. of Math. Biophys., 1945, 7, 89-93.

8. McCulloch, W. S., Finality and form, Fifteenth James Arthus Lecture, New York Academy of Science, May 2, 1946.

9. McCulloch, W. S., Machines that think and want, Lecture at the American Psychological Association, September 9, 1947.

10. McCulloch, W. S., Through the den of the metaphysician, Lecture at the University of Virginia, March 23, 1948.

11. McCulloch, W. S., Teleological mechanisms, Ann. New York Acad. Sci., 1948, 50, 4.

12. McCulloch, W. S., and Lettvin, J. Y., Somatic functions of the central nervous system, Ann. Rev. Phystol., 1948, 10, 117-132.

13. McCulloch, W. S., and Pitts, W., How we know universals. Bull. Math. Biophys., 1947, 9, 127-147.

14. McCulloch, W. S., and Pitts, W., The statistical organization of nervous aetivity, J. Amer. Statistical Assoc., 1948, 4, 91-99.

15. Rosenblueth, A., Wiener, N., and Bigelow, J., Behavior, purpose, and teleology, Philos. of Science, 1943, 10, 18-24.

16. Wiener, N., Cybernetics, New York: John Wiley and Sons, 1948.

17. Wittgenstein, L., Tractus Logico-Philosophicus, London: Paul, 1922.

DISCUSSION

DR. LORENTE DE NÓ: The main question in our minds is whether the theory as a whole is going to stand or not. I think that probably many of the details will not stand, but that the main concept will certainly remain. I'm quite sure that all of my colleagues will agree that Dr. McCulloch has brought what we know of both the anatomy and the physiology of the brain closer to an integrated whole than it has ever been before, and I want to congratulate Dr. McCulloch very much and very sincerely.

DR. VON NEUMANN: I would not like to attempt a detailed discussion of the very beautiful and very interesting presentation made by Dr. McCulloch, perhaps something like that can be done in the general discussion. I will, however, ask two questions, both dealing with only one aspect of the matter. You have emphasized that you are giving *sufficient* mechanisms and that it is in conflict with your entire philosophy at this time to claim that these are necessarily the ones that are used. You give proofs of possibility. There is, nevertheless, one point where the question of the actual mechanism is especially burning, and that is the question of memory. You have pointed out that there are positive feedbacks – reverberating circuits – built out of switching organs which are quite adequate as memory. If there were nothing else in the world except neurons, you could build memory out of neurons. My own feeling is that if one were really to construct in this way a nervous system with its known attributes it would probably take more neurons than there are, but this is an aside. My real question is this: First of all, I have observed that all neurologists seem very certain that the reverberating circuit trick is not used in making the actual memory. Amorphous intuition points in the same direction. In surmising this is not so, I have always had a bad conscience. I am not sure why they are so positive. What is the best evidence one can give for this?

The second question is this. Most neurologists with whom I have had an opportunity to talk seem to be equally convinced that memory is due to some lasting changes somewhere on the body of the nerve cell, somehow connected with alterations of thresholds. Is it not better to say that there probably is a memory organ somewhere, but that we are absolutely ignorant as to where it is – probably as ignorant as the Greeks, who located the whole intelligence in the diaphragm?

DR. MC CULLOCH: I'm afraid my answer is necessarily a bit lengthy. In the first place, I would like to contrast as sharply as possible the maximum length of what I consider reverberative memory, with the enduring memories which we bring on from childhood. I have seen a man over 80 years of age walk into a meeting of a Board of Directors and for 8 hours work out from scratch all of the details necessary for the sale of a complete railroad. He pushed the other men so as to get every piece of evidence on the table. His judgment was remarkably solid. The amount of detail involved in the transaction was enormous, and it actually took over 6 hours to get all of the requisite details on the map. He summarized that detail at the end of the meeting, in a period of a half an hour, very brilliantly, and when he came out he sat down, answered two letters that were on his desk, turned to his secretary, and said, "I have a feeling that I should have gone to a Board of Directors Meeting." He was not then, or at any later time, able to recall one iota of that meeting, and he was in that state for nearly a year before he died. This is the picture of what we call "presbyophrenia." In that state, whatever our memory organ is, we are unable to make any new record in it. Actually, the recent paths begin to fade, leaving only earlier memories to pop up. So, at that period, one is more likely to remember in detail and individually, the things that happened in childhood rather than the things of later years. Such a memory goes, quits, stops, the minute the brain is used for something else, or the minute that it comes to rest. Here is a span of at least 8 hours of high cerebral activity in carrying the details from the first moment of the meeting to the end of the meeting.

Per contra, not all memory can be of this reverberative kind. It is obvious that, although this kind of memory is carried reverberatively in the brain, it cannot endure during very deep sleep and it cannot endure during narcosis. It goes out, when the brain has a seizure and it goes out in sleep. In the one case, it goes out because the whole apparatus is pervaded by what I will call shock waves which go through it and through it and through it in the fit. In the other case, it goes out because it has no signals travelling – the brain is "shut down." I believe that only lower mechanisms are really busy in deep sleep. Now then, why do we want to attribute the memory to the brain at all? Why may it not be in the spleen or somewhere else? The answer is because injuries of the brain, but not injuries of other things, do result in losses of memory, and that is the fundamental reason for pinning it on the brain.

The next question is: Why does one attribute this, the enduring memory, to a growth process – change with use – somewhere in the excitability in neurons rather than elsewhere? Well, first, because it is a relatively lively process, and when things are growing, one tries to pin it on growth processes. Second, because it has the peculiarity that what we learn later is only a modification of what is already laid down. It is an accumulative affair of this sort. Why attribute it to the junctions of cells? Because there is where we imagine the switching takes place, and this is the kind of evidence on which we base it.

Let me tell a tale out of school, even if future evidence fails to support it. I will ask Dr. Lettvin's forgiveness later. The theory and experiments are his, although I have done some of them with him. His theory, and experiments, are designed to meet the requirements of the conditioned-reflexologists, Pavlov et al. In deference to them we will name one source of afferents U, the unconditioned afferents, which can excite an efferent R, the responsive motoneurons, and a second set of afferents C, the conditioned afferents, and – Heinrich Klüver and Warren McCulloch to the contrary notwithstanding – we will forget for the moment "stimulus equivalence" or "universals secured by averaging over groups of transformation" and treat U, C, and R as individual neurons. U can always fire R; but C can become able to do so reflexologically only if C and U are excited so as to be active concurrently. I mean that if both are concurrently active then, thereafter, C alone shall be able to fire R. Now the gist of Dr. Lettvin's analysis is that it still further simplifies the required assumption. He asks, for what do we need U – except to excite R – so why not make the simple assumption that if C and R are simultaneously active, C shall become able to fire R? Naturally this simplification should not occur to a psychologist for he has to use U to excite R. But a physiologist may put his electrodes directly upon R, if he can get them there, or he may fire it antidromically.

Now it is clearly established by the surgically and electrically perfect experiments of Donald Marquis and Arthur Ward that the intact spinal cord cannot be conditioned. But Culler and Shurrager did sometimes obtain conditioning of what they believed to be the two-neuron reflex arc, and this in experiments on by no means their technically best preparations. Wiener's theory that the spinal cord has suffered a "Wärmetod" of information by the time we are able to walk (that is, its connections are all soldered in) would account for this discrepancy. One has only to suppose that one must destroy something ending on a motoneuron to leave root room for another afferent; and the cord, no longer intact, could be conditioned even as Cajal supposed, by something else making connection with the motoneuron.

Moreover, the technically impeccable experiments of David Lloyd have proved that of two muscle antagonists at a single joint, each by its afferents (from stretch receptors) inhibits, and only inhibits, its antagonist. They show further that this inhibition occurs at the synapse on the motoneuron without time for internuncial intervention.

It follows, that if one were to cut the dorsal roots of the nerves for extension at the knee, there would be root room on its motoneurons for afferents from the flexors of the knee to get a greater hold on these motoneurons. And Dr. Lettvin's ingenious theory of synaptic transmission predicts that, if this happens, instead of inhibiting the motoneuron, these afferents will then excite them.

This, in substance, is what we did. We cut the dorsal roots of the extension reflex, stimulated its ventral roots antidromically, and at the same time stimulated the flexor muscle nerve. This we did to both for a long time, minutes or even hours, at about forty per second, from separate stimulators. Thereafter, but not before, threshold stimulation of the flexor nerve elicited contractions of the extensor. Thus Dr. Lettvin has proved that the cord, no longer intact, can be conditioned, but not quite as the psychologist would have it – for he stimulated C and R, not C and U, concurrently. He has not as yet published, and will not publish, these flndings to physiologists until his records of the times of these impulses at dorsal and ventral roots show conclusively whether or not this functionally new path is monosynaptic.

If we are not misled by the sensitivity of denervated structures, and the cord does so learn, this is crucial to psychologists. The cord is a sufficiently simple structure and is sufficiently well known for us to hope that an anatomist, with some new technique, may be able to find structural changes.

From what I have said, it should be clear that I do not think learning normally occurs in the spinal cord. Even in the earthworm learning seems to reside, albeit not in the most anterior segments, still in the forward ganglia. In mammals it may be in the midbrain, or even in the cortex, but our chances of locating anatomical changes there are negligible.

DR. VON NEUMANN: The experiment which you described – if it were done, and if the time relations were clear – would be very convincing. In this case one could at least feel certain that the "conditioning" consists in a physical change in the cell actually under consideration.

The "reverberating circuit" model for the memory does not strike me, as I said before, as a particularly elegant one. Nevertheless, it is important to know whether it is a possible model or not. I understood you to state that it is not. What is the decisive argument against it? I understood it to be that there are states when one can be fairly certain that the cortex is totally inactive and yet memory persists. What exactly is the evidence for this "total inactivity" of the cortex? Is it that one has not so far succeeded in picking up any electrical signals from it?

DR. MC CULLOCH: That's right. When our amplifier is turned up maximally, we pick up activity only from the respiratory mechanism and similar structures. Only they keep on going in deep sleep and in the coma following seizures.

DR. VON NEUMANN: Is this reliable enough to know that there is nothing else there?

DR. MC CULLOCH: Well, let's take the more powerful case – that in which you have a seizure with tremendous waves of signals through the works.

DR. VON NEUMANN: Does one know that they are really going through all channels?

DR. MC CULLOCH: I think, pretty firmly, yes; I don't believe any part of the nervous system is unaffected.

DR. VON NEUMANN: The organism is fairly well set up to protect certain parts, is it not?

DR. MC CULLOCH: The grand mal convulsion is the occasion on which that protection breaks down.

DR. VON NEUMANN: It does not seem to for memory.

DR. LORENTE DE NÓ: The difficulty in making memory reverberating paths is chiefly this: To maintain a steady state in any kind of reverberating path, the closed chains of neurons are arranged so that either you have incremental activity or decremental activity. Either the thing begins spreading to involve more and more and more neurons, or it decrements, after coming to a maximum, and then decays and disappears. Probably the secrets that Dr. McCulloch just gave us can be compared with the gramophone record. While we are playing a gramophone record, we have an articulation of impulses; later, when the record is over, memory is deposited in a different manner. As you listen to what I am saying now there are a lot of circuits operating according to those principles; but, then, when in a moment I stop talking, those signals will have stopped and memory will be somewhere else, in some other area.

DR. VON NEUMANN: I see the plausibility of what you say, but I still have a residue of uncertainty left. Your arguments about electrical circuit analogies are plausible, but they are nevertheless influenced by our particular kind of experience in this field. Your judgment based on anatomical experience is perhaps more cogent. It may be anatomically established that closed (and hence potentially reverberating) neural pathways do not exist in the necessary, vast numbers.

Another comment I would like to make is this. I see an argument that one might make against the view that memory in any form actually resides in the neurons. It is a negative argument, and far from cogent. How reasonable is it? This is the argument: There is a good deal of evidence that memory is static, unerasable, resulting from an irreversible change. (This is of course the very opposite of a "reverberating," dynamic, erasable memory.) Isn't there some physical evidence for this? If this is correct, then no memory, once acquired, can be truly forgotten. Once a memory-storage place is occupied, it is occupied forever, the memory capacity that it represents is lost; it will never be possible to store anything else there. What appears as forgetting is then not true forgetting, but merely the removal of that particular memory-storage region from a condition of rapid and easy availability to one of lower availability. It is not like the destruction of a system of files, but rather like the removal of a filing cabinet into the cellar. Indeed, this process in many cases seems to be reversible. Various situations may bring the "filing cabinet" up from the "cellar" and make it rapidly and easily available again. There are many examples of this: the "forgetting" and subsequent "remembering" or recovering of languages, telephone numbers, names – paralleling the decreased or increased need for their use.

This organizational situation is a very plausible one, if there is a memory which is much larger than the available switching facilities for its selective use. Indeed, if the memory is thus larger than its switching system, it will be necessary to introduce a system of priorities for various parts of the memory. Each part may then, upon occasion, be moved into regions with rapid accessibility, or into regions with less rapid accessibility. Or, rather, it may not be moved from region to region, but be connected to quickly or to less quickly functioning portions of the available switching system.

If this is so, then the memory cannot reside in the actual switching organs in the neurons, and its capacity must be much greater than that represented by the switching system. One must then postulate a very high-capacity memory organ or organization, with considerable bottlenecks at the "input" and "output," that is, at the points of contact represented by the switching system.

Does this sound plausible, or is there some flaw in my argument?

DR. BROSIN: May I break here with the tradition of immediate reply? This is a very large subject, and I would like to see if there are other commentators. You may gather the evidence, and if we do not have time enough today, you can have a full dress performance tomorrow.

DR. GERARD: I would like to ask a few questions, some of which have already been touched upon. I have a very trivial one first. I didn't quite see why you place such disparate emphasis on the manipulation of the output of the brain in efferent systems and paths, as compared to the problem of the manipulation of the input of the brain. If I correctly understood you, you are not particularly worried about the input side. I don't see why that is, and I would like to have you explain it a little more clearly.

I have, also, two other points that touch on this memory problem. If learning and remembering are based on growth processes of some sort, then they should not be basically different from developmental and maturational behavior; and yet it seems to me that some of the most striking experimental work in the past does emphasize a very fundamental difference between the maturational learning in the nervous system and acquiring a new behavioral capacity – experiential acquiring of new behavior possibilities. The former takes place certainly without any external experience, but you can see that there is internal experience. On the question of rest and activity of the nervous system, several members of the audience, during our intermission, raised the question with me whether you are not neglecting something that you might call automatic activity of neurons. The assumption is that the output will be determined by the input, rather than by something happening independently of the input. I will put the question to you in this way. Do I correctly assume that you were suggesting that the scansion machine in your projection area mechanism is the spontaneous brain wave, and whether it starts there or below is immaterial? I would like to have that elaborated and made a little bit clearer. I personally am surprised at the answer you gave Dr. von Neumann, that the brain can be completely quiet. I don't believe that electrically, or in any other way, it is ever completely quiet at any time except in death.

DR. MC CULLOCH: I said, except for lower mechanisms.

DR. GERARD: I believe that, even in the other mechanisms, I have never seen a completely silent brain.

DR. MC CULLOCH: No, I don't believe that brain matter is ever completely quiet. I'll take care of that question later. For the moment it is enough that there are times when no signals are reverberating.

DR. GERARD: If there is a separate memory organ, along the line of Dr. von Neumann's comment, in which you have your files easily accessible or down in the basement, that would argue against the memory traces being associated with the neurons themselves throughout the brain. What about the reversible amnesia problem, where all past memory vanishes for long periods and then comes back again? If learning involves the establishment of new functionally effective connections between neurons in the brain (whether by growth, by physiological change in threshold, or what not), and if that depends on activation of neurons and association with experience, then it seems to me that it should follow that if the threshold of neurons is held low, just in general during the experiencing of experience, learning should be enhanced. Dr. Lettvin raised the level of excitability of neurons in the nervous system. There is more chance of a particular input leaving a permanent modification, or even a temporary one. I hope some of the psychologists here can bring the evidence in, but I don't know. However, as far as I'm aware, conditioning under the influence of stimulating drugs has not changed the rate of this conditioning.

The last question I should really leave for Dr. Lashley, since it is in his field. If these networks of neurons (even allowing for considerable interchangeability of particular elements of the net) are organized so beautifully in the striate and elsewhere for these particular functions, then how do you account for some of Dr. Lashley's critical experiments on destruction of different parts of the brain and the retention of learning, memory, and all the rest of it?

DR. KÖHLER: I admire the courage with which Dr. McCulloch tries to relate his neurophysiology to facts in psychology. But, when in a skeptical mood, I sometimes feel like criticizing the results. Take the example of visual shapes which, as we all know, are generally recognized in a peripheral position (or in a larger size), even if, heretofore, they have been seen only in foveal projection (or in a smaller size). Dr. McCulloch's explanation of such achievements introduces more histological assumptions ad hoc than seem compatible with usual standards of plausibility. In fact, he does not seem himself to maintain that a real brain functions in this fashion. Why then the elaborate constructions? Most probably the reason is that the atomistic character of Dr. McCulloch's neurophysiology prevents any direct approach to relationally determined facts such as visual shapes. The difficulty seems to be strongly felt, and special sets of neuron connections are now being constructed which merely serve to remove the difficulties caused by the main atomistic premise. Would it not be simpler never to make this atomistic assumption? If we think of cortical function in terms of continuous field physics rather than of impulses in neurons, the difficulty never arises. The contours of retinal images are projected upon the visual cortex by nerve impulses. Let us assume that here they constitute the boundary conditions of field processes such as electric currents. Under these circumstances, there will be for each set of boundary conditions, that is, for each shape, a particular distribution of a directly interrelated function; in other words, each shape will be cortically represented by a specific process. If the characteristics of such a process remain approximately constant, independently of its location and size, then recognition of a shape in a new place or size offers no problem which is not also present in the recognition of a color in a new place.

Incidentally, it seems to me misleading to assume that the present problem is mainly a problem of recognition, and therefore of memory.

When two objects are given simultaneously in different places while the eyes do not move, we can compare these objects, and say whether they have the same shape. Once more the implication is that visual shapes are associated with specific processes.

Occasionally, I am afraid, Dr. McCulloch uses psychological terms in a strangely diluted sense. In fact, sometimes little is left of what they actually mean in psychology. But the change is never mentioned. People will therefore tend to believe that, when such terms are now being related to neurophysiological hypotheses, it is their real psychological contents which are given a physiological interpretation. They will not notice that the essential characteristics of the facts in question are tacitly being ignored. I have an uneasy feeling that this may happen even to the theorist himself. Thus Dr. McCulloch likes to call a nerve impulse a "proposition." Moreover, he says that the occurrence of a given nerve impulse "implies" the occurrence of preceding impulses (in other neurons), by which the given impulse has been started. But, typically, a proposition is concerned with a relation between certain terms, whatever the relation may be in individual instances. A cortical situation would therefore correspond to a proposition if in this situation the cortical counterparts of two terms were functionally related in one specific fashion or another. A nerve impulse does not in this sense relate two terms to each other. At least in Dr. McCulloch's neurophysiology, a nerve impulse seems to be a particularly lonely event. How, then, can a nerve impulse represent a proposition? Some discussions of nerve impulses and of their equivalence to facts in psychology make me feel that, inadvertently, an extremely learned histologist and neurophysiologist is tacitly supposed to watch the human brain continually, and that this expert always knows how impulses must be interpreted in psychological terms. He probably tells the owner of the brain what psychological facts he must have when impulses travel in this or that part of the cortical machine. For without this help, what could induce a person to think of a specific proposition, that is, a particular relation between two terms, when an impulse travels in a certain fiber? Since, actually, no such expert is available, the characteristic forms of the various psychological facts must be directly given by the functional characteristics of corresponding cortical processes. But, to repeat, if this is the case, it cannot be nerve impulses which give propositions their relational character. For they have no such character themselves. For the same reason, there can be no connection between the psychological experience that one fact implies another fact and the behavior of a nerve impulse. A present impulse implies, say, preceding impulses in other neurons (McCulloch's example) only in the mind of a neurophysiologist who knows what must have happened a moment ago at a certain synapse. As the present impulse travels along its fiber, it knows nothing of preceding impulses.

For a moment, I must come back to a criticism to which I have referred once before. It must be a hard task to give psychological facts interpretations in terms of nerve impulses. For when this task arises, and is apparently accepted, the theorists soon forget what they must now be expected to do, and turn to other problems which are only indirectly connected with the original problems. Invariably, such substitute problems are more accessible to explanations in terms of nerve impulses. On the other hand, since they are somehow related to the problems which were actually to be solved, psychological concepts which are essential in the latter will naturally also be mentioned when the substitutes are being discussed. Thus, if interpretations in terms of nerve impulses seem to work in the case of the substitutes, both the theorists and others will easily believe that actually the original problems have been solved. For this is what the theorists had promised to achieve.

Take "having a goal" as an example. Before we realize what is happening, the task of explaining this psychological fact in terms of nerve impulses has been replaced by another task: Once a person has a goal, how is the goal actually reached? Naturally, if this is done by overt action, both centrifugal and centripetal nerve impulses will play an important role in the process. It is also a most sensible suggestion that the action is steered in the right direction by negative feedback. But do we learn in this fashion what "having a goal" is in terms of nerve impulses? Plainly, we do not. Nonetheless, we may be so strongly impressed by what seems to have been achieved that we forget what had to be achieved. Of course, it must be difficult to understand "having a goal" as a matter of nerve impulses. "Having a goal" is again a relational situation. When a person has a goal, his self (in a purely empirical sense) is dynamically related to a certain object, and therefore, probably; the neural counterpart of the self to the counterpart of the object. Moreover, the nature of the relation depends entirely upon the perceived characteristics of the object and the state of the self at the given time. The theorists themselves seem to doubt whether interrelations of this kind can be mediated by nerve impulses, which are described as atomic events par excellence. Otherwise, why should the theorists prefer to discuss something else, namely, goal-directed action? And yet, "having a goal" is a problem which must be handled quite apart from overt action in reaching the goal. For people often have goals while they do not yet know how these goals can possibly be reached. It also seems probable that a really adequate interpretation of "reaching a goal" presupposes a correct interpretation of "having a goal." As a goal is being reached, the dynamic relation between the self and the goal, which seems to represent a store of energy, is gradually being changed-until eventually, when the goal has been reached, this energy is spent. I have a suspicion that the negative feedback involved in the change refers to the store of energy implicit in "having a goal." But, of course, this again is thinking in terms of field physics.

I will remark only in passing that the substitution of one problem for another occurs also in Dr. McCulloch's treatment of "value." He does not give us a theory of value in terms of nerve impulses. But values may conflict, just as many other things may conflict, and then the question arises which value will win in a given conflict; that is, which is the stronger value. It is this question with which he prefers to deal. But since the same question may be asked with regard to many facts which are not at all values, we have obviously once more lost our way. We may easily believe that we are actually dealing with the problem of value as such; and this belief will be strengthened by the fact that, in formulating our new problem, we may still mention the concept "value." But the problem what value means in terms of nerve impulses has in the meantime been forgotten.

Quite probably, Dr. McCulloch will not be impressed by these arguments. He may feel that I am accepting certain premises of which he does not approve. First of all, he is likely to say that the structural characteristics of cortical processes need not agree with the structures of corresponding psychological facts. Actually, he has just told us that the hypothetical cortical counterpart of an idea must fulfill only one condition: It must always occur when the idea occurs. More specifically, he has once said that the cortical counterpart of a square may be a "four-spoked form, not at all like a square." I cannot agree with this statement for the following reason. It would not be difficult to give subjects a series of tests in which they would have to respond to one structural characteristic of a square after another in overt action. Under these circumstances, the form of their actions would directly follow from the corresponding structural properties of the square. Their actions would prove, for instance, that their square has four straight sides, that pairs of these sides are parallel, that the angles have all the same size, and so forth. From the point of view of natural science, how can this happen if the cortical counterpart of the square has no corresponding characteristics? Does Dr. McCulloch suggest that such characteristics exist only in the square as an "apparition" (his term), that is, as a mental fact, and that, quite apart from the cortical situation, this mental fact as such determines what the subjects are doing? There is an old name for this view. It is called dualism. I find it hard to believe that dualism appeals to Dr. McCulloch. But in this connection he does argue as though he were a dualist.

If we consider how the visual square (the apparition) comes into existence, we meet with the same difficulty. How can a cortical process such as that of a square give rise to an apparition with certain structural characteristics, if these characteristics are not present in the process itself? According to Dr. McCulloch, this is actually the case. But if we follow the example of physics, we shall hesitate to accept his view. In physics, the structural characteristics of a state of affairs are given by the structural properties of the factors which determine that state of affairs. The magnetic field around a long conductor with circular cross section obviously describes circles; the electrostatic field around a charged sphere is symmetrical with regard to the center of the sphere, and so forth. Situations in physics which depend upon the spatial distribution of given conditions never have more, and more specific, structural characteristics than are contained in the conditions. To be sure, this rule holds only so long as the medium in which a physical situation develops is homogeneous, that is, devoid of special conditions of its own. For instance, the field around a charged sphere will no longer be symmetrical about the center if the environment contains various dielectrics in an arbitrary arrangement.

If we apply this lesson to the way in which the cortical counterpart of a square gives rise to this square, we must choose between two possibilities. Either the structural characteristics of the visual square are fully determined by its cortical counterpart. Then this cortical process must have the structural characteristics of the square. Or we assume that the visual square has structural characteristics of its own which are not present in the cortical process. Then the world of apparitions, the psychological world, constitutes a particular medium with special determining conditions, quite apart from cortical conditions; and it is these conditions in the mental world which add the structural characteristics not contained in the cortical process. The second alternative is, of course, again tantamount to dualism. It seems that if we do not want to be dualists we must accept psychophysical isomorphism.

DR. BROSIN: May I again beg your indulgence and give you a full opportunity to continue later? Dr. Lashley.

DR. LASHLEY: I am very much in sympathy with the type of development represented in the last two papers. I think any understanding of the nervous system we may acquire must be developed within the framework of our knowledge of the activities of the individual neuron. There may be additional factors introduced by combinations of which we know little or nothing at present, but the general principles seem to me to be fundamentally correct. At the present time, however, such a formulation involves a very great oversimplification of the problems. The behavior which is explained is behavior which never occurs in the intact organism. It is an hypothetical behavior derived from the assumptions of the system rather than a description of observed phenomena. A visual object maintains its continuity in spite of constant fluctuations in the position of the eyes and shifts in its position on the retina. By a series of special assumptions concerning neural organization this phenomenon of stimulus equivalence can be accounted for in terms of impulse switching. But in recognition of the visual object it makes little difference whether the image of the whole object falls upon the retina. If only part is seen at any one time, the entire form is rapidly reconstructed from the series of images of parts. The temporal sequence of part figures is combined with the spatial orientation of eye movements to give spatial continuity to the whole. This phenomenon requires a new set of assumptions to make the theory of impulse switching applicable. I somewhat question the utility of a theory which has to be revised to fit each special case.

In its present form the theory of impulse switching involves, I believe, assumptions concerning the accuracy and uniformity of neuronic structure which are not justified by the facts. We have been studying individual variations in the number and arrangement of neurons in the cerebral cortex. We find a wide range of individual differences in cell number and size in corresponding areas of different brains of animals which are grossly indistinguishable in behavior. Two brains may differ by as much as 50 per cent in the number of neurons in the temporal lobe or by 100 per cent in the average size of cells in the superior frontal convolution. A given type of cell may be present or absent from the auditory cortex or operculum. Yet the fundamental behavioral activities of these animals are the same. The anatomic variability is so great as to preclude, I believe, any theory which assumes regularity and precision of anatomic arrangement. Such facts lead me to believe that theories of neuron interaction must be couched, not in terms of the activity of individual cells but in terms of mass relations among the cells. Even the simplest bit of behavior requires the integrated action of millions of neurons; the activity of any single neuron can have little influence on the whole, just as the path of an individual molecule of a gas has little influence on the gas pressure. It is questionable whether specific instances of behavior can ever be dealt with in terms of the activity of individual neurons; the complexity is too great. We shall probably have to use a different kind of model, a model which can be explained in principle by individual neuron action but which involves a somewhat different set of concepts and laws of action. These laws may eventually be derived from study of the individual neuron when those properties are directly observed. At present, however, many of the properties ascribed to the individual neuron are inferred from the activities of neuron masses, and explanations based on such inferred properties are circular and, perhaps, spurious.

Some of the specific hypotheses which have been formulated by Dr. McCulloch seem to me to meet with serious difficulties. He has suggested a reverberatory system between the striate cortex, the suppressor band of area 19, the prestriate region, and the thalamus. I have just removed the prestriate region (including areas 18 and 19) from a series of monkeys and also the frontal eye fields (another suppressor area) singly and in combination. I have been able to detect no visual disturbances whatever following the operations. In no case have we been able to detect significant perceptual disturbances after removal of suppressor areas or of supposed sensory associative areas in monkeys. The specific hypotheses which Dr. McCulloch has suggested for the action of the visual and auditory analyzers imply a definite spatial position of the analyzing mechanisms. Experimentally they are not there.

This leads to the general problem raised by most of the experimental studies of effects of cerebral lesions. Limited lesions or interruptions of transcortical connections produce few or no symptoms. Behavior seems not to depend upon any localized conducting pathways within the cortex. Habits are not stored in any limited area. Such facts point to the conclusion that there is multiple representation of every function. I see no other way of meeting this difficulty except by assuming some sort of reduplicated network of equivalent functional circuits. In other words, we cannot deal with individual conditioned reflex arcs but only with a multiplicity of interacting circuits whose excitatory effects can be transmitted around various types of cortical interruption.

One other point, in relation to the problems of memory raised by Dr. von Neumann: For memory there is the same problem of equivalence as for transneural conduction. I have found, for example, that one sixtieth of the visual cortex of the rat will mediate visual memories and it may be any sixtieth, provided it includes part of the central projection field. Here, again, there must be some sort of multiple representation. The memory is not stored in a single locus.

Now consider the nature of a memory. It is not a single item which can be filed in a single neuron or reverberatory circuit. It is always the capacity to reproduce a series of events, to reproduce a complex sensory pattern or a series of motor activities. Such neural events involve the activity of millions of cells. I have come to believe that almost every nerve cell in the cerebral cortex may be excited in every activity. I shall give some quantitative evidence of this tomorrow. Differential behavior is determined by the combinations of cells acting together rather than by cells which participate only in particular bits of behavior. The same neurons which maintain the memory traces and participate in the revival of a memory are also involved, in different combinations, in thousands of other memories and acts. The memory trace is the capacity of many neurons to work together in certain permutations. In a system of interconnected neurons the number of possible permutations may greatly exceed the number of switching mechanisms. Perhaps this answers Dr. von Neumann's difficulty with regard to number of elements. It is also an argument against the dynamic as opposed to static character of the memory trace.

DR. BROSIN: Dr. Weiss, have you anything to add?

DR. WEISS: Much of my comment had better be left for a later part of the symposium. For the present I want to point out that we are actually dealing here with two different problems. Namely, first a statistical consideration, as it were, of whether or not the number of elements present in the nervous system and their various interrelations is sufficient to account for the number and variety of things it can do. We realize, and this gives us intellectual comfort, that the number of possible constellations is large enough to allow for the observed variety of behavior. This statement, to quote McCulloch, would be merely tautological. As a biologist, I am more interested in the second problem, and that is the precise pathways and chains of processes through which, out of the infinite variety of possibilities, just the appropriate sequence and selection are activated which lead to a given appropriate organized response. And if we deal with these mechanisms not as abstract categories, but in concrete terms, then I see some serious and realistic difficulties arising for any theory of nervous networks that requires the amount of precision postulated in the schemes here presented. The study of the developed nervous system, with which the anatomists, physiologists, and psychologists are usually working, suggests a high degree of precision in the arrangement of the constituent elements, but it must be realized that this impression is illusory. The organizational stability of a performance of the nervous system is much greater than the precision of the underlying structural apparatus.

I have referred above to the persistence of the response after experimental or pathological interference with the anatomical substrata of nervous activity, but want to point now to an even more impressive fact, namely, the great variability in the degree of precision of the anatomical networks in the course of development. The fact is that we frequently suspect a given neuronal precision setup as being relevant for a particular neural function, but often find that in an earlier stage of development this function will be performed in essentially the same way without that particular structural precision scheme having even developed as yet. In general, many a condition which we would think essential from the study of the developed nervous system loses pertinence when studies of earlier stages show that things work very much the same even in its absence. This must be emphasized particularly in connection with the present discussion of the relation between input and output of the nervous system. It is a fact that most of the basic motor patterns of behavior are developed within the nervous system by virtue of the laws of its own embryonic differentiation without the aid of, and prior to the appearance of, a sensory input from the outside world. The basic configuration of the motor patterns, therefore, cannot possibly be a direct product of the patterns of the sensory input. A study of the development of the nervous system and of behavior forces us to consider the output of the nervous system and its patterns as primarily preformed within the nervous system and ready for use, requiring the sensory input for release, facilitation, and modification, but not for its primary shaping.

This brings us to the fundamental alternative, to which I think Dr. Gerard has likewise referred, of whether the central nervous system is merely a clearing house for input-to-output messages, or whether it generates activities of its own and has patterns of activities of its own, the elements of which are not pieced together by, and reflections of, the sensory input. Dr. Köhler has likewise touched on this fundamental difference in the interpretation of the realities of the nervous system. No theory of the nervous system can claim to represent the facts if it ignores the central autonomy of the basic patterns of motor performance. This autonomy impresses us not only in the studies of the development, but also in studies on reconstitution after injury of the nervous system, which touches on a question Dr. von Neumann has raised with regard to learning. This is the question of whether learning implies a complete reorganization of the nervous network with a resetting of relations among individual neurons, or the acquisition of a new performance, which will merely supersede, rather than replace, the older performance. This question can be crucially studied by disarranging the peripheral motor apparatus by crossing tendons or nerves, and thus rendering the original impulse patterns inadequate for the performance of a given act. Experiments by my former student, Sperry, have shown that rats cannot relearn their motor coordination to meet such new situations. Studies we have made on patients with transplanted tendons after partial infantile paralysis show that they can learn to use the transplanted muscle in its new function, but precise electro-myographic records show that the muscle is apt to lapse back into its innate phase of activity, thus proving that the learning act does not dissolve the original patterns of motor organization. Evidence of this kind demonstrates clearly that the act of learning does not consist of merely a recombination of individual neuronal elements. On the basis of all existing evidence, the nervous system must not be conceived of as a network of monotonic elements, but as a hierarchical system in which groups of neuronal complexes of different kinds are acting as units, the properties of which determine the configuration of the output pattern. Some of these higher units are rigidly fixed in their functions, others are modifiable by experience. I fail to see this hierarchical principle duly reflected in the theory of a monotonic network of units such as has been discussed in this session.

* * *

The text was originally edited and rendered into PDF file for the e-journal <www.vordenker.de> by *E. von Goldammer*

vordenker
ISSN 1619-9324

How to cite:
Warren St. McCulloch: Why the Mind Is in the Brain, in: www.vordenker.de (Edition: Winter 2008/09), J. Paul (Ed.), URL: < http://www.vordenker.de/ggphilosophy/mcculloch_why-the-mind-is-in-the-head.pdf > — published in: "Warren S. McCulloch: Embodiments of Mind", The MIT Press, 1965, p.72-87. (first published in: Cerebral Mechanisms in Behavior, L.A. Jeffress (ed.), The Hixon Symposium, Jon Wiley Publ., New York 1951, pp. 42-57)