A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY

WARREN S. MCCULLOCH and WALTER H. PITTS

Because of the "all-or-none" character of nervous activity, neural events and the relations among them can be treated by means of propositional logic. It is found that the behavior of every net can be described in these terms, with the addition of more complicated logical means for nets containing circles; and that for any logical expression satisfying certain conditions, one can find a net behaving in the fashion it describes. It is shown that many particular choices among possible neurophysiological assumptions are equivalent, in the sense that for every net behaving under one assumption, there exists another net which behaves under the other and gives the same results, although perhaps not in the same time. Various applications of the calculus are discussed.

INTRODUCTION

THEORETICAL neurophysiology rests on certain cardinal assumptions. The nervous system is a net of neurons, each having a soma and an axon. Their adjunctions, or synapses, are always between the axon of one neuron and the soma of another. At any instant a neuron has some threshold, which excitation must exceed to initiate an impulse. This, except for the fact and the time of its occurrence, is determined by the neuron, not by the excitation. From the point of excitation the impulse is propagated to all parts of the neuron. The velocity along the axon varies directly with its diameter, from less than one meter per second in thin axons, which are usually short, to more than 150 meters per second in thick axons, which are usually long. The time for axonal conduction is consequently of little importance in determining the time

originally published in: Bulletin of Mathematical Biophysics, Vol. 5, 1943, p. 115-133

of arrival of impulses at points unequally remote from the same source. Excitation across synapses occurs predominantly from axonal terminations to somata. It is still a moot point whether this depends upon irreciprocity of individual synapses or merely upon prevalent anatomical configurations. To suppose the latter requires no hypothesis *ad hoc* and explains known exceptions, but any assumption as to cause is compatible with the calculus to come. No case is known in which excitation through a single synapse has elicited a nervous impulse in any neuron, whereas any neuron may be excited by impulses arriving at a sufficient number of neighboring synapses within the period of latent addition, which lasts less than one quarter of a millisecond. Observed temporal summation of impulses at greater intervals is impossible for single neurons and empirically depends upon structural properties of the net. Between the arrival of impulses upon a neuron and its own propagated impulse there is a synaptic delay of more than half a millisecond. During the first part of the nervous impulse the neuron is absolutely refractory to any stimulation. Thereafter 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. Frequent activity augments this subnormality. Such specificity as is possessed by nervous impulses depends solely upon their time and place and not on any other specificity of nervous energies. Of late only inhibition has been seriously adduced to contravene this thesis. Inhibition is the termination or prevention of the activity of one group of neurons by concurrent or antecedent activity of a second group. Until recently this could be explained on the supposition that previous activity of neurons of the second group might so raise the thresholds of internuncial neurons that they could no longer be excited by neurons of the first group, whereas the impulses of the first group must sum with the impulses of these internuncials to excite the now inhibited neurons. Today, some inhibitions have been shown to consume less than one millisecond. This excludes internuncials and requires synapses through which impulses inhibit that neuron which is being stimulated by impulses through other synapses. As yet experiment has not shown whether the refractoriness is relative or absolute. We will assume the latter and demonstrate that the difference is immaterial to our argument. Either variety of refractoriness can be accounted for in either of two ways. The "inhibitory synapse" may be of such a kind as to produce a substance which raises the threshold of the neuron, or it may be so placed that the local disturbance produced by its excitation opposes the alteration induced by the otherwise excitatory synapses. Inasmuch as position is already known to have such effects in the case of electrical stimulation, the first hypothesis is to be excluded unless and until it be substantiated, for the second involves no new hypothesis. We have, then, two explanations of inhibition based on the same general premises, differing only in the assumed nervous nets and, consequently, in the time required for inhibition. Hereafter we shall refer to such nervous nets as *equivalent in the extended sense*. Since we are concerned with properties of nets which are invariant under equivalence, we may make the physical assumptions which are most convenient for the calculus.

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. Two difficulties appeared. The first concerns facilitation and extinction, in which antecedent activity temporarily alters responsiveness to subsequent stimulation of one and the same part of the net. The second concerns learning, in which activities concurrent at some previous time have altered the net permanently, so that a stimulus which would previously have been inadequate is now adequate. But for nets undergoing both alterations, we can substitute equivalent fictitious nets composed of neurons whose connections and thresholds are unaltered. But one point must be made clear: neither of us conceives the formal equivalence to be a factual explanation. *Per contra!*—we regard facilitation and extinction as dependent upon continuous changes in threshold related to electrical and chemical variables, such as after-potentials and ionic concentrations; and learning as an enduring change which can survive sleep, anaesthesia, convulsions and coma. The importance of the formal equivalence lies in this: that the alterations actually underlying facilitation, extinction and learning in no way affect the conclusions which follow from the formal treatment of the activity of nervous nets, and the relations of the corresponding propositions remain those of the logic of propositions.

The nervous system contains many circular paths, whose activity so regenerates the excitation of any participant neuron that reference to time past becomes indefinite, although it still implies that afferent activity has realized one of a certain class of configurations over time. Precise specification of these implications by means of recursive functions, and determination of those that can be embodied in the activity of nervous nets, completes the theory.

THE THEORY: NETS WITHOUT CIRCLES

We shall make the following physical assumptions for our calculus.

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 on 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.

To present the theory, the most appropriate symbolism is that of Language II of R. Carnap (1938), augmented with various notations drawn from B. Russell and A. N. Whitehead (1927), including the *Principia* conventions for dots. Typographical necessity, however, will compel us to use the upright 'E' for the existential operator instead of the inverted, and an arrow ('→') for implication instead of the horseshoe. We shall also use the Carnap syntactical notations, but print them in boldface rather than German type; and we shall introduce a functor S, whose value for a property P is the property which holds of a number when P holds of its predecessor; it is defined by '$S(P)\ (t) \,.\equiv.\, P(Kx)\,.\,t = x$')'; the brackets around its argument will often be omitted, in which case this is understood to be the nearest predicate-expression [$\boldsymbol{Pr}$] on the right. Moreover, we shall write $S^2\boldsymbol{Pr}$ for $S(S(\boldsymbol{Pr}))$, etc.

The neurons of a given net $\mathfrak{N}$ may be assigned designations 'c_1', 'c_2', . . . , 'c_n'. This done, we shall denote the property of a number, that a neuron c_i fires at a time which is that number of synaptic delays from the origin of time, by 'N' with the numeral i as subscript, so that $N_i(t)$ asserts that c_i fires at the time t. N_i is called the *action* of c_i. We shall sometimes regard the subscripted numeral of 'N' as if it belonged to the object-language, and were in a place for a functoral argument, so that it might be replaced by a number-variable [z] and quantified; this enables us to abbreviate long but finite disjunctions and conjunctions by the use of an operator. We shall employ this locution quite generally for sequences of $\boldsymbol{Pr}$; it may be secured formally by an obvious disjunctive definition. The predicates 'N_1', 'N_2', . . . , comprise the syntactical class '$\boldsymbol{N}$'.

Let us define the *peripheral afferents* of $\mathfrak{N}$ as the neurons of $\mathfrak{N}$ with no axons synapsing upon them. Let N_1, . . . , N_p denote the actions of such neurons and N_{p+1}, N_{p+2}, . . . , N_n those of the rest. Then a *solution of* $\mathfrak{N}$ will be a class of sentences of the form $\boldsymbol{S}_i$: $N_{p+1}\ (z_1) \,.\equiv.\, \boldsymbol{Pr}_i\ (N_1, N_2, \ldots, N_p, z_1)$, where $\boldsymbol{Pr}_i$ contains no free variable save z_1 and no descriptive symbols save the $\boldsymbol{N}$ in the argument [$\boldsymbol{Arg}$], and possibly some constant sentences [$\boldsymbol{sa}$]; and such that each $\boldsymbol{S}_i$ is true of $\mathfrak{N}$. Conversely, given a $\boldsymbol{Pr}_1\ ({}^1p^1{}_1, {}^1p^1{}_2, \ldots, {}^1p^1{}_p, z_1, \boldsymbol{s})$, containing no free variable save those in its $\boldsymbol{Arg}$, we shall say that it is *realizable in the narrow sense* if there exists a net $\mathfrak{N}$

and a series of N_i in it such that $N_1(z_1) \,.\!\equiv\!.\, Pr_1(N_1, N_2, \ldots, z_1, sa_1)$ is true of it, where sa_1 has the form $N(0)$. We shall call it *realizable in the extended sense*, or simply *realizable*, if for some n $S^n(Pr_1)(p_1, \ldots, p_p, z_1, s)$ is realizable in the above sense. c_{pi} is here the realizing neuron. We shall say of two laws of nervous excitation which are such that every S which is realizable in either sense upon one supposition is also realizable, perhaps by a different net, upon the other, that they are equivalent assumptions, in that sense.

The following theorems about realizability all refer to the extended sense. In some cases, sharper theorems about narrow realizability can be obtained; but in addition to greater complication in statement this were of little practical value, since our present neurophysiological knowledge determines the law of excitation only to extended equivalence, and the more precise theorems differ according to which possible assumption we make. Our less precise theorems, however, are invariant under equivalence, and are still sufficient for all purposes in which the exact time for impulses to pass through the whole net is not crucial.

Our central problems may now be stated exactly: first, to find an effective method of obtaining a set of computable S constituting a solution of any given net; and second, to characterize the class of realizable S in an effective fashion. Materially stated, the problems are to calculate the behavior of any net, and to find a net which will behave in a specified way, when such a net exists.

A net will be called *cyclic* if it contains a circle: i.e., if there exists a chain $c_i, c_{i+1}, \ldots$ of neurons on it, each member of the chain synapsing upon the next, with the same beginning and end. If a set of its neurons $c_1, c_2, \ldots, c_p$ is such that its removal from $\mathfrak{N}$ leaves it without circles, and no smaller class of neurons has this property, the set is called a *cyclic* set, and its cardinality is the *order of* $\mathfrak{N}$. In an important sense, as we shall see, the order of a net is an index of the complexity of its behavior. In particular, nets of zero order have especially simple properties; we shall discuss them first.

Let us define a *temporal propositional expression* (a *TPE*), designating a *temboral propositional function* (*TPF*), by the following recursion:

1. A ${}^1p^1\,[z_1]$ is a *TPE*, where p_1 is a predicate-variable.

2. If S_1 and S_2 are *TPE* containing the same free individual variable, so are SS_1, $S_1 v S_2$, $S_1 . S_2$ and $S_i . \sim S_2$.

3. Nothing else is a *TPE*.

Theorem I

Every net of order 0 *can be solved in terms of temporal propositional expressions.*

Let c_i be any neuron of $\mathfrak{N}$ with a threshold $\theta_i > 0$, and let c_{i1}, c_{i2}, ... , c_{ip} have respectively n_{i1}, n_{i2}, ... , n_{ip} excitatory synapses upon it. Let c_{j1}, c_{j2}, ... , c_{jq} have inhibitory synapses upon it. Let κ_i be the set of the subclasses of $\{n_{i1}, n_{i2}, \ldots, n_{ip}\}$ such that the sum of their members exceeds θ_i. We shall then be able to write, in accordance with the assumptions mentioned above,

$$N_i(z_1) \;.\equiv.\; S\left\{\prod_{m=1}^{q} \sim N_{jm}(z_1) \;.\sum_{\alpha\epsilon K_i}\;\prod_{s\epsilon\alpha} N_{is}(z_1)\right\} \qquad (1)$$

where the '$\sum$' and '$\prod$' are syntactical symbols for disjunctions and conjunctions which are finite in each case. Since an expression of this form can be written for each c_i which is not a peripheral afferent, we can, by substituting the corresponding expression in (1) for each N_{jm} or N_{is} whose neuron is not a peripheral afferent, and repeating the process on the result, ultimately come to an expression for N_i in terms solely of peripherally afferent N, since $\mathfrak{N}$ is without circles. Moreover, this expression will be a *TPE*, since obviously (1) is; and it follows immediately from the definition that the result of substituting a *TPE* for a constituent $p(z)$ in a *TPE* is also one.

Theorem II

Every TPE *is realizable by a net of order zero.*

The functor S obviously commutes with disjunction, conjunction, and negation. It is obvious that the result of substituting any S_i, realizable in the narrow sense (i.n.s.), for the $p(z)$ in a realizable expression S_1 is itself realizable i.n.s.; one constructs the realizing net by replacing the peripheral afferents in the net for S_1 by the realizing neurons in the nets for the S_i. The one neuron net realizes $p_1(z_1)$ i.n.s., and Figure 1-a shows a net that realizes $Sp_1(z_1)$ and hence SS_2, i.n.s., if S_2 can be realized i.n.s. Now if S_2 and S_3 are realizable then $S^m S_2$ and $S^n S_3$ are realizable i.n.s., for suitable m and n. Hence so are $S^{m+n} S_2$ and $S^{m+n} S_3$. Now the nets of Figures 1b, c and d respectively realize $S(p_1(z_1) \vee p_2(z_1))$, $S(p_1(z_1) \,.\, p_2(z_1))$, and $S(p_1(z_1) \,.\, \sim p_2(z_1))$ i.n.s. Hence $S^{m+n+1}(S_1 \vee S_2)$, $S^{m+n+1}(S_1 \,.\, S_2)$, and $S^{m+n+1}(S_1 \,.\, \sim S_2)$ are realizable i.n.s. Therefore $S_1 \vee S_2 S_1 \,.\, S_2 S_1 \,.\, \sim S_2$ are realizable if S_1 and S_2 are. By complete induction, all *TPE* are realizable. In this way all nets may be regarded as built out of the fundamental elements of Figures 1a, b, c, d, precisely as the temporal propositional expressions are generated out of the operations of precession, disjunction, conjunction, and conjoined negation. In particular, corresponding to any description of state, or distribution of the values *true* and *false* for the actions of all the neurons of a net save that which makes them all false, a single neuron is constructible whose firing is a necessary and sufficient condition for the validity of that description. Moreover, there is always an indefinite number of topologically different nets realizing any *TPE*.

Theorem III

Let there be given a complex sentence $\mathbf{S}_1$ *built up in any manner out of elementary sentences of the form* $\mathbf{p}(\mathbf{z}_1 - \mathbf{zz})$ *where* **zz** *is any numeral, by any of the propositional connections: negation, disjunction, conjunction, implication, and equivalence. Then* $\mathbf{S}_1$ *is a* TPE *and only if it is false when its constituent* $\mathbf{p}(\mathbf{z}_1 - \mathbf{zz})$ *are all assumed false—i.e., replaced by false sentences—or that the last line in its truth-table contains an* '**F**',*—or there is no term in its Hilbert disjunctive normal form composed exclusively of negated terms.*

These latter three conditions are of course equivalent (Hilbert and Ackermann, 1938). We see by induction that the first of them is necessary, since $p(z_1 - zz)$ becomes false when it is replaced by a false sentence, and $S_1 \vee S_2$, $S_1 \,.\, S_2$ and $S_1 \,.\, \sim S_2$ are all false if both their constituents are. We see that the last condition is sufficient by remarking that a disjunction is a *TPE* when its constituents are, and that any term

$$S_1 \,.\, S_2 \,.\, \ldots \,.\, S_m \,.\, \sim S_{m+1} \,.\, \sim \,.\, \ldots \,.\, \sim S_n$$

can be written as

$$(S_1 \,.\, S_2 \,.\, .\, .\, .\, S_m) \,.\, \sim (S_{m+1} v\, S_{m+2} v \,.\, .\, .\, .\, v\, S_n),$$

which is clearly a *TPE.*

The method of the last theorems does in fact provide a very convenient and workable procedure for constructing nervous nets to order, for those cases where there is no reference to events indefinitely far in the past in the specification of the conditions. By way of example, we may consider the case of heat produced by a transient cooling.

If a cold object is held to the skin for a moment and removed, a sensation of heat will be felt; if it is applied for a longer time, the sensation will be only of cold, with no preliminary warmth, however transient. It is known that one cutaneous receptor is affected by heat, and another by cold. If we let N_1 and N_2 be the actions of the respective receptors and N_3 and N_4 of neurons whose activity implies a sensation of heat and cold, our requirements may be written as

$$N_3(t) \,:\, \equiv \,:\, N_1(t-1) \,.\, \mathrm{v} \,.\, N_2(t-3) \,.\, \sim N_2(t-2)$$
$$N_4(t) \,.\, \equiv \,.\, N_2(t-2) \,.\, N_2(t-1)$$

where we suppose for simplicity that the required persistence in the sensation of cold is, say, two synaptic delays, compared with one for that of heat. These conditions clearly fall under Theorem III. A net may consequently be constructed to realize them, by the method of Theorem II. We begin by writing them in a fashion which exhibits them as built out of their constituents by the operations realized in Figures 1a, b, c, d: i.e., in the form

$$N_3(t) \,.\, \equiv \,.\, S\{N_1(t) \,\mathrm{v}\, S[(SN_2(t)) \,.\, \sim N_2(t)]\}$$
$$N_4(t) \,.\, \equiv \,.\, S\{[SN_2(t)] \,.\, N_2(t)\}.$$

First we construct a net for the function enclosed in the greatest number of brackets and proceed outward; in this case we run a net of the form shown in Figure 1a from c_2 to some neuron c_a, say, so that

$$N_a(t) \,.\, \equiv \,.\, SN_2(t).$$

Next introduce two nets of the forms 1c and 1d, both running from c_a and c_2, and ending respectively at c_4 and say c_b. Then

$$N_4(t) \,.\, \equiv \,.\, S[N_a(t) \,.\, N_2(t)] \,.\, \equiv \,.\, S[(SN_2(t)) \,.\, N_2(t)].$$
$$N_b(t) \,.\, \equiv \,.\, S[N_a(t) \,.\, \sim N_2(t)] \,.\, \equiv \,.\, S[(SN_2(t)) \,.\, \sim N_2(t)].$$

Finally, run a net of the form 1b from c_1 and c_b to c_3, and derive

$$N_3(t) \,.\equiv.\, S[N_1(t) \lor N_b(t)]$$
$$.\equiv.\, S\{N_1(t) \lor S[(SN_2(t)) \,.\, \sim N_2(t)]\}.$$

These expressions for $N_3(t)$ and $N_4(t)$ are the ones desired; and the realizing net *in toto* is shown in Figure 1e.

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. The same illusion, of course, could also have been produced under various other assumptions about the behavior of the cutaneous receptors, with correspondingly different nets.

We shall now consider some theorems of equivalence: i.e., theorems which demonstrate the essential identity, save for time, of various alternative laws of nervous excitation. Let us first discuss the case of *relative inhibition.* By this we mean the supposition that the firing of an inhibitory synapse does not absolutely prevent the firing of the neuron, but merely raises its threshold, so that a greater number of excitatory synapses must fire concurrently to fire it than would otherwise be needed. We may suppose, losing no generality, that the increase in threshold is unity for the firing of each such synapse; we then have the theorem:

Theorem IV

Relative and absolute inhibition are equivalent in the extended sense.

We may write out a law of nervous excitation after the fashion of (1), but employing the assumption of relative inhibition instead; inspection then shows that this expression is a *TPE*. An example of the replacement of relative inhibition by absolute is given by Figure 1f. The reverse replacement is even easier; we give the inhibitory axons afferent to c_i any sufficiently large number of inhibitory synapses apiece.

Second, we consider the case of extinction. We may write this in the form of a variation in the threshold θ_i after the neuron c_i has fired; to the nearest integer—and only to this approximation is the variation in threshold significant in natural forms of excitation—this may be written as a sequence $\theta_i + b_j$ for j synaptic delays after firing, where $b_j = 0$ for j large enough, say $j = M$ or greater. We may then state

Theorem V

Extinction is equivalent to absolute inhibition.

For, assuming relative inhibition to hold for the moment, we need merely run M circuits $\mathcal{T}_1$, $\mathcal{T}_2$, ... $\mathcal{T}_M$ containing respectively 1, 2, ... , M neurons, such that the firing of each link in any is sufficient to fire the next, from the neuron c_i back to it, where the end of the circuit $\mathcal{T}_j$ has just b_j inhibitory synapses upon c_i. It is evident that this will produce the desired results. The reverse substitution may be accomplished by the diagram of Figure 1g. From the transitivity of replacement, we infer the theorem. To this group of theorems also belongs the well-known

Theorem VI

Facilitation and temporal summation may be replaced by spatial summation.

This is obvious: one need merely introduce a suitable sequence of delaying chains, of increasing numbers of synapses, between the exciting cell and the neuron whereon temporal summation is desired to hold. The assumption of spatial summation will then give the required results. See e.g. Figure 1h. This procedure had application in showing that the observed temporal summation in gross nets does not imply such a mechanism in the interaction of individual neurons.

The phenomena of learning, which are of a character persisting over most physiological changes in nervous activity, seem to require the possibility of permanent alterations in the structure of nets. The simplest such alteration is the formation of new synapses or equivalent local depressions of threshold. We suppose that some axonal terminations cannot at first excite the succeeding neuron; but if at any time the neuron fires, and the axonal terminations are simultaneously excited, they become synapses of the ordinary kind, henceforth capable of exciting the neuron. The loss of an inhibitory synapse gives an entirely equivalent result. We shall then have

Theorem VII

Alterable synapses can be replaced by circles.

This is accomplished by the method of Figure 1i. It is also to be remarked that a neuron which becomes and remains spontaneously active can likewise be replaced by a circle, which is set into activity by a peripheral afferent when the activity commences, and inhibited by one when it ceases.

THE THEORY: NETS WITH CIRCLES

The treatment of nets which do not satisfy our previous assumption of freedom from circles is very much more difficult than that case. This is largely a consequence of the possibility that activity may be set up in a circuit and continue reverberating around it for an indefinite period of time, so that the realizable ***Pr*** may involve reference to past events of an indefinite degree of remoteness. Consider such a net $\mathfrak{N}$, say of order p, and let $c_1, c_2, \ldots, c_p$ be a cyclic set of neurons of $\mathfrak{N}$. It is first of all clear from the definition that every $\boldsymbol{N_s}$ of $\mathfrak{N}$ can be expressed as a *TPE*, of $N_1, N_2, \ldots, N_p$ and the absolute afferents; the solution of $\mathfrak{N}$ involves then only the determination of expressions for the cyclic set. This done, we shall derive a set of expressions [***A***]:

$$N_i(z_1) \;.\equiv.\; \boldsymbol{Pr}_i[S^{n_{i1}} N_1(z_1), S^{n_{i2}} N_2(z_1), \ldots, S^{n_{ip}} N_p(z_1)], \tag{2}$$

where $\boldsymbol{Pr}_i$ also involves peripheral afferents. Now if n is the least common multiple of the n_{ij}, we shall, by substituting their equivalents according to (2) in (3) for the N_j, and repeating this process often enough on the result, obtain $\boldsymbol{S}$ of the form

$$N_i(z_1) \;.\equiv.\; \boldsymbol{Pr}_1[S^n N_1(z_1), S^n N_2(z_1), \ldots, S^n N_p(z_1)]. \tag{3}$$

These expressions may be written in the Hilbert disjunctive normal form as

$$N_i(z_1) \;.\equiv.\; \sum_{\substack{\alpha\varepsilon k \\ \beta_\alpha \varepsilon k}} S_a \prod_{j\varepsilon k} S^n N_j(z_1) \prod_{j\varepsilon\beta_\alpha} \sim S^n N_j(z_1), \text{ for suitable } \kappa, \tag{4}$$

where $\boldsymbol{S_a}$ is a *TPE* of the absolute afferents of $\mathfrak{N}$. There exist some 2^p different sentences formed out of the pN_i by conjoining to the conjunction of some set of them the conjunction of the negations of the rest. Denumerating these by $X_1(z_1)$, $X_2(z_1)$, ..., $X_{2p}(z_1)$, we may, by use of the expressions (4), arrive at an equipollent set of equations of the form

$$X_i(z_1) \;.\equiv.\; \sum_{j=1}^{2p} Pr_{ij}(z_1) \;.\; S^n X_j(z_1). \qquad (5)$$

Now we import the subscripted numerals i,j into the object-language: i.e., define Pr_1 and Pr_2 such that $Pr_1(zz_1,z_1) \;.\equiv.\; X_i(z_1)$ and $Pr_2(zz_1,zz_2,z_1) \;.\equiv.\; Pr_{ij}(z_1)$ are provable whenever zz_1 and zz_2 denote i and j respectively.
Then we may rewrite (5) as

$$\begin{aligned}&(z_1)zz_p : Pr_1(z_1, z_3)\\ &\quad .\equiv.\; (Ez_2)zz_p \,.\, Pr_2(z_1, z_2, z_3 - zz_n) \,.\, Pr_1(z_2, z_3 - zz_n)\end{aligned} \qquad (6)$$

where zz_m denotes n and zz_p denotes 2^p. By repeated substitution we arrive at an expression

$$\begin{aligned}&(z_1)zz_p : Pr_1(z_1, zz_n\, zz_2) \;.\equiv.\; (Ez_2)zz_p\; (Ez_3)zz_p \ldots (Ez_n)zz_p.\\ &Pr_2(z_1, z_2. zz_n\,(zz_2 - 1)) \,.\, Pr_2(z_2, z_3, zz_n\,(zz_2 - 1)) \ldots\ldots\end{aligned} \qquad (7)$$

$Pr_2(z_{n-1},z_n,0) \,.\, Pr_1(z_n,0)$, for any numeral zz_2 which denotes s.

This is easily shown by induction to be equipollent to

$$\begin{aligned}&(z_1)zz_p :.\; Pr_1(z_1, zz_n\, zz_2) :\equiv: (Ef)\; (z_2)\, zz_2 - 1\, f(z_2\, zz_n)\\ &\leqq zz_p \,.\, f(zz_n\, zz_2) = z_1 \,.\, Pr_2(f(zz_n\,(z_2 + 1)),\\ &f(zz_n\, z_2)) \,.\, Pr_1(f(0), 0)\end{aligned} \qquad (8)$$

and since this is the case for all zz_2, it is also true that

$$\begin{aligned}&(z_4)\; (z_1)zz_p : Pr_1(z_1, z_4) \;.\equiv.\; (Ef)\; (z_2)\; (z_4 - 1) \,.\, f(z_2)\\ &\leqq zz_p \,.\, f(z_4) = z_1 f(z_4) = z_1 \,.\, Pr_2[f(z_2 + 1), f(z_2), z_2] \,.\\ &Pr_1[f(\mathrm{res}\;(z_4, zz_n)), \mathrm{res}\;(z_4, zz_n)],\end{aligned} \qquad (9)$$

where zz_n denotes n, res (r,s) is the residue of r mod s and zz_p denotes 2^p. This may be written in a less exact way as

$$\begin{aligned}N_i(t) \;.\equiv.\; (E\phi)\; (x)t - 1 \,.\, \phi(x) \leqq 2^p \,.\, \phi(t) = i \,.\\ P[\phi(x+1), \phi(x) \,.\, N_{\phi(0)}\; (0)],\end{aligned}$$

where x and t are also assumed divisible by n, and Pr_2 denotes P. From the preceding remarks we shall have

Theorem VIII

> *The expression* (9) *for neurons of the cyclic set of a net* $\mathfrak{N}$ *together with certain* TPE *expressing the actions of other neurons in terms of them, constitute a solution of* $\mathfrak{N}$.

Consider now the question of the realizability of a set of $\boldsymbol{S}_i$. A first necessary condition, demonstrable by an easy induction, is that

$$(z_2)z_1 \,.\, p_1(z_2) \equiv p_2(z_2) \,.\rightarrow.\, \boldsymbol{S}_i \equiv \boldsymbol{S}_i \begin{Bmatrix} p_1 \\ p_2 \end{Bmatrix} \tag{10}$$

should be true, with similar statements for the other free p in $\boldsymbol{S}_i$: i.e., no nervous net can take account of future peripheral afferents. Any $\boldsymbol{S}_i$ satisfying this requirement can be replaced by an equipollent $\boldsymbol{S}$ of the form

$$\begin{aligned} &(Ef)\ (z_2)z_1\ (z_3)zz_p : f \epsilon\, Pr_{mi} \\ &\qquad : f(z_1, z_2, z_3\ = 1 \,.\equiv.\, p_{z3}(z_2) \end{aligned} \tag{11}$$

where zz_p denotes p, by defining

$$\begin{aligned} Pr_{mi} = \hat{f}[(z_1)\ (z_2)z_1(z_3)zz_p :.\, f(z_1, z_2, z_3) = 0 \,.\mathbf{v}.\, f(z_1, z_2, z_3) \\ = 1 : f(z_1, z_2, z_3) = 1 \,.\equiv.\, p_{z3}(z_2) :\rightarrow: \boldsymbol{S}_i]. \end{aligned}$$

Consider now these series of classes α_i, for which

$$\begin{aligned} N_i(t) :\equiv: (E\phi)\ (x)t(m)q : \phi\epsilon\alpha_i : N_m(x) \,.\equiv.\, \phi(t, x, m) = 1. \\ [i = q + 1, \cdots, M] \end{aligned} \tag{12}$$

holds for some net. These will be called *prehensible* classes. Let us define the *Boolean ring* generated by a class of classes κ as the aggregate of the classes which can be formed from members of κ by repeated application of the logical operations; i.e., we put

$$\begin{aligned} \mathcal{R}(\kappa) = p\text{‘}\hat{\lambda}[(\alpha, \beta) : \alpha\epsilon\kappa \\ \rightarrow \alpha\epsilon\lambda : \alpha, \beta\epsilon\lambda \,.\rightarrow.\, -\alpha, \alpha \,.\, \beta, \alpha \,\mathbf{v}\, \beta\epsilon\lambda]. \end{aligned}$$

We shall also define

$$\begin{aligned} &\overline{\mathcal{R}}(\kappa) \,.=.\, \mathfrak{R}(\kappa) - \iota\text{‘}p\text{‘} - \text{‘‘}\kappa, \\ &\mathcal{R}_e(\kappa) \ = p\text{‘}\,\hat{\lambda}[(\alpha, \beta) : \alpha\epsilon\kappa \rightarrow \alpha\epsilon\lambda \,.\rightarrow.\, -\alpha, \alpha \,.\, \beta, \alpha \,\mathbf{v}\, \beta, S\,\text{‘‘}\alpha\epsilon\hat{\lambda} \\ &\overline{\mathcal{R}}_e(\kappa) \ = \mathcal{R}_e(\kappa) - \iota\text{‘}p\text{‘} - \text{‘‘}\kappa, \end{aligned}$$

and

$$\sigma(\psi, t) = \hat{\phi}[(m) \,.\, \phi(t + 1, t, m) = \psi(m)].$$

The class $\mathscr{R}_e(\kappa)$ is formed from κ in analogy with $\mathscr{R}(\kappa)$, but by repeated application not only of the logical operations but also of that which replaces a class of properties $P \epsilon \alpha$ by $S(P) \epsilon S``\alpha$. We shall then have the

Lemma

$\boldsymbol{Pr}_1(p_1, p_2, \ldots , p_m, z_1)$ is a TPE if and only if

$$(z_1)\ (p_1, \ldots , p_m)\ (Ep_{m+1}) : p_{m+1} \epsilon \overline{\mathscr{R}}_e(\{p_1, p_2, \ldots , p_m\}) \\ p_{m+1}(z_1) \equiv \boldsymbol{Pr}_1(p_1, p_2, \ldots , p_m, z_1) \tag{13}$$

is true; and it is a TPE not involving 'S' if and only if this holds when '$\overline{\mathscr{R}}_c$' is replaced by '$\overline{\mathscr{R}}$', and we then obtain

Theorem IX

A series of classes α_1, α_2, $\ldots$ α_s is a series of prehensible classes if and only if

$$(Em)\ (En)\ (p)n(i)\ (\psi) :. (x)m\psi(x) = 0 \vee \psi(x\ \ = 1 : \rightarrow : (E\beta) \\ (Ey)m \,.\, \psi(y) = 0 \,.\, \beta\epsilon\mathscr{R}[\hat{\gamma}((Ei) \,.\, \gamma = \alpha_i)) \,.\, \mathrm{v} \,.\, (x)m \,. \\ \psi(x) = 0 \,.\, \beta\epsilon\overline{\mathscr{R}}[\hat{\gamma}((E_i) \,.\, \gamma = \alpha_i)] : (t)\ (\phi) : \phi\epsilon\alpha_i \,. \\ \sigma(\phi, nt + p) \,.\rightarrow.\, (Ef) \,.\, f\epsilon\beta \,.\, (w)m(x)t - 1 \,. \\ \phi(n(t + 1) + p, nx + p, w) = f(nt + p, nx + p, w). \tag{14}$$

The proof here follows directly from the lemma. The condition is necessary, since every net for which an expression of the form (4) can be written obviously verifies it, the ψ's being the characteristic functions of the $\boldsymbol{S}_a$ and the β for each ψ being the class whose designation has the form $\prod\limits_{i\epsilon\alpha} \boldsymbol{Pr}_i \prod\limits_{j\epsilon\beta} \boldsymbol{Pr}_j$, where $\boldsymbol{Pr}_k$ denotes α_k for all k. Conversely, we may write an expression of the form (4) for a net $\mathfrak{N}$ fulfilling prehensible classes satisfying (14) by putting for the $\boldsymbol{Pr}_a$ $\boldsymbol{Pr}$ denoting the ψ's and a $\boldsymbol{Pr}$, written in the analogue for classes of the disjunctive normal form, and denoting the α corresponding to that ψ, conjoined to it. Since every $\boldsymbol{S}$ of the form (4) is clearly realizable, we have the theorem.

It is of some interest to consider the extent to which we can by knowledge of the present determine the whole past of various special nets: i.e., when we may construct a net the firing of the cyclic set of whose neurons requires the peripheral afferents to have had a set of past values specified by given functions ϕ_i. In this case the classes α_i of the last theorem reduced to unit classes; and the condition may be transformed into

$$\begin{aligned}&(E\,m, n)\ (p)n(i, \psi)\ (Ej) : .\ (x)m : \psi(x) = 0\ .\ \mathrm{v}\ .\ \psi(x) = 1 : \\ &\quad \phi_i \epsilon \sigma(\psi, nt + p) : \rightarrow : (w)m(x)t - 1\ .\ \phi_i(n(t + 1) \\ &\quad + p, nx + p, w) = \phi_j(nt + p, nx + p, w) : . \\ &\quad (u, v)\ (w)m\ .\ \phi_i(n(u + 1) + p, nu + p, w) \\ &\quad = \phi_i(n(v + 1) + p, nv + p, w).\end{aligned}$$

On account of limitations of space, we have presented the above argument very sketchily; we propose to expand it and certain of its implications in a further publication.

The condition of the last theorem is fairly simple in principle, though not in detail; its application to practical cases would, however, require the exploration of some 2^{2n} classes of functions, namely the members of $\mathcal{R}(\{\alpha_1, \ldots, \alpha_s\})$. Since each of these is a possible β of Theorem IX, this result cannot be sharpened. But we may obtain a sufficient condition for the realizability of an $\boldsymbol{S}$ which is very easily applicable and probably covers most practical purposes. This is given by

Theorem X

Let us define a set of $\boldsymbol{K}$ of $\boldsymbol{S}$ by the following recursion:

1. Any *TPE* and any *TPE* whose arguments have been replaced by members of $\boldsymbol{K}$ belong to $\boldsymbol{K}$;

2. If $\boldsymbol{Pr}_1(z_1)$ is a member of $\boldsymbol{K}$, then $(z_2)z_1\ .\ \boldsymbol{Pr}_1(z_2)$, $(\mathrm{E}z_2)z_1$ $\boldsymbol{Pr}_1(z_2)$, and $\boldsymbol{C}_{mn}(z_1)\ .\ \boldsymbol{s}$ belong to it, where $\boldsymbol{C}_{mn}$ denotes the property of being congruent to m modulo n, $m < n$.

3. *The set* $\boldsymbol{K}$ *has no further members.*

Then every member of $\boldsymbol{K}$ is realizable.

For, if $\boldsymbol{Pr}_1(z_1)$ is realizable, nervous nets for which

$$\begin{aligned}N_i(z_1)\ .\equiv .\ \boldsymbol{Pr}_1(z_1)\ .\ SN_i(z_1)\\ N_i(z_1)\ .\equiv .\ \boldsymbol{Pr}_1(z_1)\ \mathrm{v}\ SN_i(z_1)\end{aligned}$$

are the expressions of equation (4), realize $(z_2)z_1\ .\ \boldsymbol{Pr}_1(z_2)$ and

$(E\, z_2)z_1 \,.\, Pr_1(z_2)$ respectively; and a simple circuit, $c_1, c_2, \ldots, c_n$, of n links, each sufficient to excite the next, gives an expression

$$N_m(z_1) \,.\, \equiv \,.\, N_1(0) \,.\, C_{mn}$$

for the last form. By induction we derive the theorem.

One more thing is to be remarked in conclusion. 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.

CONSEQUENCES

Causality, which requires description of states and a law of necessary connection 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. 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

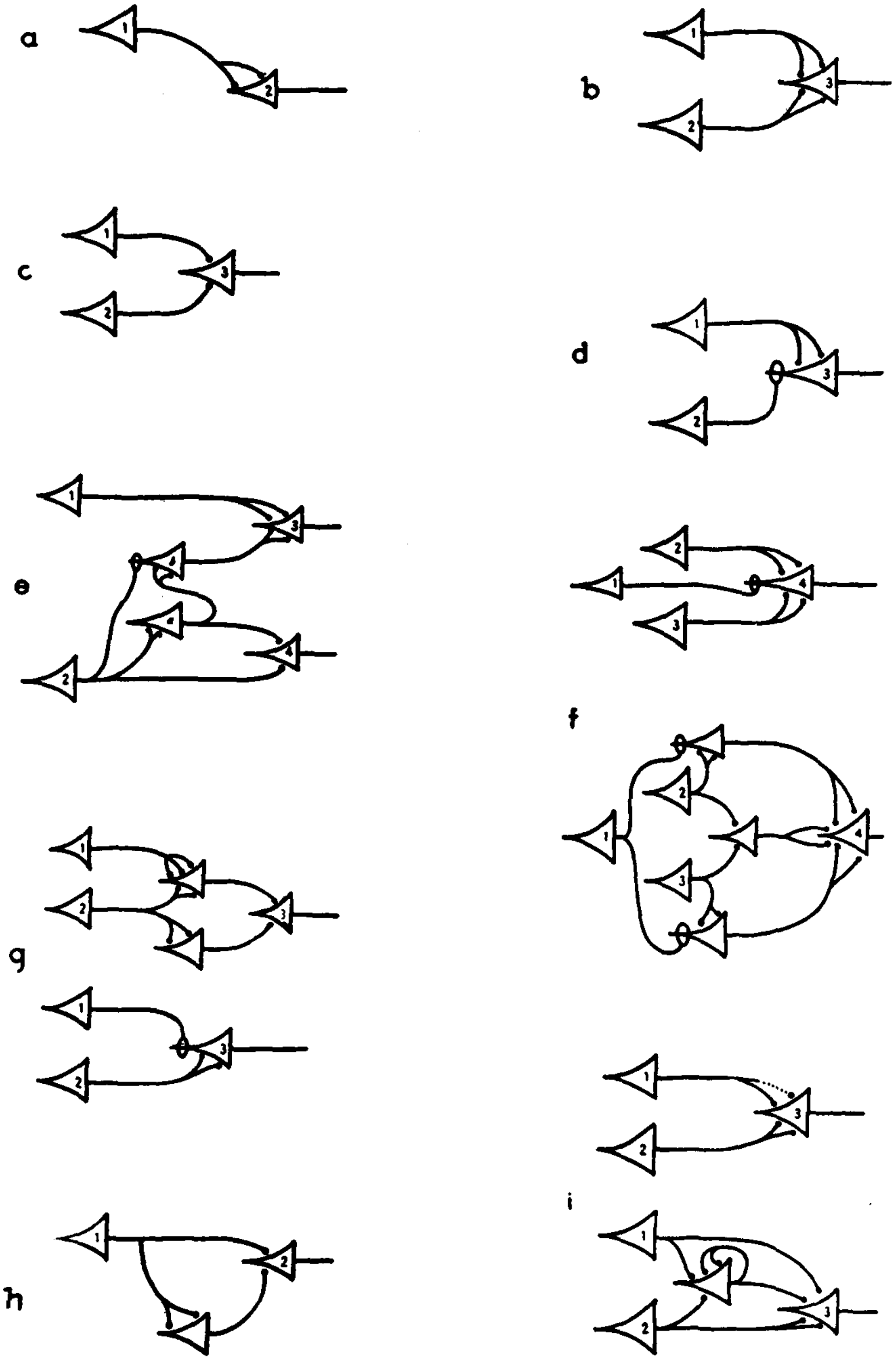

Figure 1

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.

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 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." With determination of the net, the unknowable object of knowledge, the "thing in itself," ceases to be unknowable.

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

←

EXPRESSION FOR THE FIGURES

In the figure the neuron c_i is always marked with the numeral i upon the body of the cell, and the corresponding action is denoted by 'N' with i as subscript, as in the text.

Figure 1a $N_2(t) \,.\equiv.\, N_1(t-1)$

Figure 1b $N_3(t) \,.\equiv.\, N_1(t-1) \vee N_2(t-1)$

Figure 1c $N_3(t) \,.\equiv.\, N_1(t-1) \,.\, N_2(t-1)$

Figure 1d $N_3(t) \,.\equiv.\, N_1(t-1) \,.\, \sim N_2(t-1)$

Figure 1e $N_3(t) :\equiv: N_1(t-1) \,.\vee.\, N_2(t-3) \,.\sim N_2(t-2)$
$N_4(t) \,.\equiv.\, N_2(t-2) \,.\, N_2(t-1)$

Figure 1f $N_4(t) :\equiv: \sim N_1(t-1) \,.\, N_2(t-1) \vee N_3(t-1) \,.\vee.\, N_1(\mathrm{t}-1) \,.$
$N_2(t-1) \,.\, N_3(t-1)$
$N_4(t) :\equiv: \sim N_1(t-2) \,.\, N_2(t-2) \vee N_3(t-2) \,.\vee.\, N_1(t-2) \,.$
$N_2(t-2) \,.\, N_3(t-2)$

Figure 1g $N_3(t) \,.\equiv.\, N_2(t-2) \,.\sim N_1(t-3)$

Figure 1h $N_2(t) \,.\equiv.\, N_1(t-1) \,.\, N_1(t-2)$

Figure 1i $N_3(t) :\equiv: N_2(t-1) \,.\vee.\, N_1(t-1) \,.\, (Ex)t-1 \,.\, N_1(x) \,.\, N_2(x)$

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.

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 eye-witnesses, 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 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 wont to call *mental* are rigorously deducible from present neurophysiology. 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. 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.

REFERENCES

1. Carnap, R.: *The Logical Syntax of Language*. New York, Harcourt, Brace and Company, 1938.
2. Hilbert, D., und Ackermann, W.: *Grundüge der Theoretischen Logik*. Berlin, J. Springer, 1927.
3. Whitehead, A. N., and Russell, B.: *Principia Mathematica*. Cambridge, Cambridge University Press, 1925.

originally published in: Bulletin of Mathematical Biophysics, Vol. 5, 1943, p. 115-133