S. C. Kleene. Representation of events in nerve nets and finite automata.
In C. E. Shannon and J. McCarthy, editors, Automata Studies, pages 3--41.
Princeton University Press, 1956. Annals of Mathematics Studies 34.

REPRESENTATION OF EVENTS IN NERVE NETS AND FINITE AUTOMATA*

S. C. Kleene

INTRODUCTION

1 Stimuli and Response

An organism or an automaton receives stimuli via its sensory receptor organs, and performs actions via its effector organs. To say that certain actions are a response to certain stimuli means, in the simplest case, that the actions are perfomed when and only when those stimuli occur.

In the general case both the stimuli and the actions may be very complicated.

In order to simplify the analysis, we may begin by leaving out of account the complexities of the response. We reason that any sort of stimulation, or briefly any event, which affects action in the sense that different actions ensue according as the event occurs or not, under some set of other circumstances held fixed, must have a representation in the state of the organism or automaton, after the event has occurred and prior to the ensuing action.

So we ask what kind of events are capable of being represented in the state of an automaton.

For explaining actions as responses to stimuli it would remain to study the manner in which the representations of events (a kind of internal response) lead to the overt responses.

Our principal result will be to show (in Sections 7 and 9) that all and only the events of a certain class called "regular events" are representable.

2 Nerve Nets and Behavior

McCulloch and Pitts [McC 43] in their fundamental paper on the logical analysis of nervous activity formulated certain assumptions which we shall

*The material in this article is drawn from Project RAND Research Memorandum RM-704 (15 December 1951, 101 pages) under the same title and by the author. It is used now by permission of the RAND Corporation. The author's original work on the problem was supported by the RAND Corporation during the summer of 1951.

recapitulate in Section 3.

In showing that each regular event is representable in the state of a finite automaton, the automaton we use is a McCulloch-Pitts nerve net. Thus their neurons are one example of a kind of "universal elements" for finite automata.

The McCulloch-Pitts assumptions were put forward as an abstraction from neuro-physiological data. We shall not be concerned with the question of how exactly the assumptions fit. They seem to fit roughly up to a point, though one of McCulloch's and Pitts' results is that certain alternative assumptions can explain the same kind of behavior. With increasing refinement in the neuro-physiological data the emphasis is no doubt on respects in which the assumptions do not fit.

Our theoretical objective is not dependent on the assumptions fitting exactly. It is a familiar strategem of science, when faced with a body of data too complex to be mastered as a whole, to select some limited domain of experiences, some simple situations, and to undertake to construct a model to fit these at least approximately.

Having set up such a model, the next step is to seek a thorough understanding of the model itself. It is not to be expected that all features of the model will be equally pertinent to the reality from which the model was extracted. But after understandig the model, one is in a better position to see how to modify or adapt it to fit the limited data better or to fit a wider body of data and when to seek some fundamentally different kind of explanation.

McCulloch and Pitts in their original paper give a theory for nerve nets without circles [Part II of their paper] and a theory for arbitrary nerve nets [Part III]. The present article is partly an exposition of their results; but we found the part of their paper dealing with arbitrary nerve nets obscure, so we have proceeded independently there.

Although we are concerned with the model itself rather than its application, a few remarks on the latter may prevent misunderstanding.

To take one example, as consideration of the model shows, memory can be explained on the basis of reverberating cycles of nerve impulses. This seems a plausible explanation for short-term memories. For long-term memories, it is implausible on the ground of fatigue, also on the ground that calculations on the amount of material stored in the memory would call for too many neurons [McC 49], and also on the basis of direct experimental evidence that temporary suppression of nervous activity does not cut off memory [GER 53].

The McCulloch-Pitts assumptions give a nerve net the character of a digital automaton, as contrasted to an analog mechanism in the sense familiar in connection with computing machines. Some physiological processes of

control seem to be analog. Just as in mathematics continuous processes can be approximated by discrete ones, analog mechanisms can be approximated in their effect by digital ones. Nevertheless, the analog or partly analog controls may for some purposes be the simplest and most economical.

An assumption of the present mathematical theory is that there are no errors in the functioning of neurons. Of course this is unrealistic both for living neurons and for the corresponding units of a mechanical automaton. It is the natural procedure, however, to begin with a theory of what happens assuming no malfunctioning. Indeed in our theory we may represent the occurrence of an event by the firing of a single neuron. Biologically it is implausible that important information should be represented in an organism in this way. But by suitable duplication and interlacing of circuits, one could then expect to secure the same results with small probability of failure in nets constructed of fallible neurons.

Finally, we repeat that we are investigating McCulloch-Pitts nerve nets only partly for their own sake as providing a simplified model of nervous activity, but also as an illustration of the general theory of automata, including robots, computing machines and the like. What a finite automaton can and cannot do is thought to be of some mathematical interest intrinsically, and may also contribute to better understanding of problems which arise on the practical level.

PART I. NERVE NETS

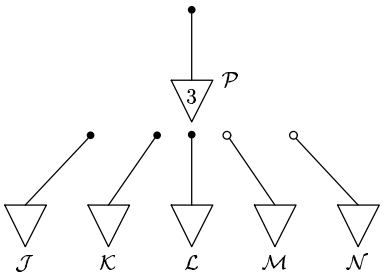
3 McCulloch-Pitts Nerve Nets

Under the assumptions of McCulloch and Pitts [McC 43], a nerve cell or *neuron* consists of a body or *soma*, whence nerve fibers (*axons*) lead to one or more *endbulbs*.

A *nerve net* is an arrangement of a finite number of neurons in which each endbulb of any neuron is adjacent to (*impinges* on) the soma of not more than one neuron (the same or another); the separating gap is a *synapse*. Each endbulb is either *excitatory* or *inhibitory* (not both).

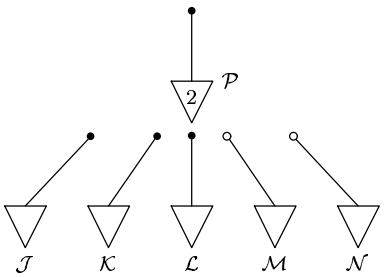
We call the neurons (zero or more) on which no endbulbs impinge *input neurons*; the others, *inner neurons*.

At equally separated moments of time (which we take as the integers on a time scale, the same for all neurons in a given net), each neuron of the net is either *firing* or *not firing* (being *quiet*). For an input neuron, the firing or not-firing at any moment t is determined by conditions outside the net. One can suppose each is impinged on by a sensory receptor organ, which



$$P(t) \equiv J(t-1)\ \&\ K(t-1)\ \&\ L(t-1)\ \&\ \overline{M(t-1)}\ \&\ \overline{N(t-1)}.$$

Figure 1: Conjunctive Net

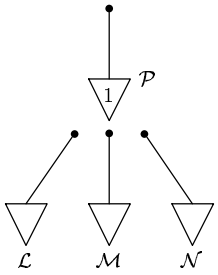


$$P(t) \equiv [[J(t-1)\ \&\ K(t-1)]\ \vee\ [J(t-1)\ \&\ L(t-1)]\ \vee\ [K(t-1)\ \&\ L(t-1)]]\ \&\ \overline{M(t-1)}\ \&\ \overline{N(t-1)}.$$

Figure 2

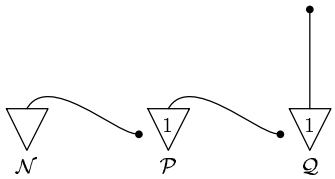
under suitable conditions in the environment causes the neuron to fire at time t. For an inner neuron, the condition for firing at time t is that at least a certain number h (the *threshold* of that neuron) of the excitatory endbulbs and none of the inhibitory endbulbs impinging on it belong to neurons which fired at time $t-1$.

For illustration, consider the nerve net shown in Figure 1, with input neurons $\mathcal{J}, \mathcal{K}, \mathcal{L}, \mathcal{M}$ and $\mathcal{N}$, and inner neuron $\mathcal{P}$. Excitatory endbulbs are shown as dots, and inhibitory as circles. The threshold of $\mathcal{P}$ is 3, as shown by the number in the triangle representing its soma. The formula written below the net expresses in logical symbolism that neuron $\mathcal{P}$ fires at time t, if and only if all of $\mathcal{J}, \mathcal{K}$ and $\mathcal{L}$ and none of $\mathcal{M}$ and $\mathcal{N}$ fired at time $t-1$. We are writting "$P(t)$" to say that neuron $\mathcal{P}$ fires at time t, "$J(t-1)$" to say that $\mathcal{J}$ fired at $t-1$, etc. The symbol "$\equiv$" means *if and only if* (or *is equivalent to*), "&" means *and*, "$\vee$" means *or* (in the non-exclusive sense), and "—" means *not*.



$$P(t) \equiv L(t-1) \ \vee \ M(t-1) \ \vee \ N(t-1).$$

Figure 3: Disjunctive Net



$$P(t) \equiv N(t-1),$$
$$Q(t) \equiv N(t-2).$$

Figure 4: Delay Net

t	$\mathcal{N}_1$	$\mathcal{N}_2$
p	1	0
$p-1$	1	1
$p-2$	0	1
	...	

Figure 5

4 The Input to a Nerve Net

Consider a nerve net with k input neurons $\mathcal{N}_1, \ldots, \mathcal{N}_k$. We assume $k \geq 1$, until 6.3. The *input* (or *experience*) over all past time up to the present moment inclusive can be described by a table with k columns corresponding to the input neurons, and with rows corresponding to the moments counting backward from the present moment p. The positions are filled with 0's and 1's, where 0 is to stand for quiescence, and 1 for firing, of the neuron in question at the moment in question.

For example with $k = 2$ the table might be like Figure 5. The 1 in the first row and first column means that $\mathcal{N}_1$ fires at time p, the 0 in the third row and first column that $\mathcal{N}_1$ did not fire at time $p-2$, etc. If this table is extended down infinitely, we have a representation of the input, thought of as extending over all past time, which for the time being we treat as infinite. (In Section 6 we shall reconsider the matter.)

By an *event* we shall mean any property of the input. In other words, any subclass of the class of all the possible tables describing the input over all past time (and ending with the present p inclusive, except when otherwise stated) constitutes an event, which *occurs* when the table describing the actual input belongs to this subclass.

Examples of events with two neurons $\mathcal{N}_1$ and $\mathcal{N}_2$ are:

(1) $\mathcal{N}_1$ fires at time p.

(2) $\mathcal{N}_2$ does not fire at time p, and $\mathcal{N}_1$ fired at time $p-1$.

(3) One of $\mathcal{N}_1$ and $\mathcal{N}_2$ fires at time p.

(4) $\mathcal{N}_1$ and $\mathcal{N}_2$ both fire at time p.

(5) $\mathcal{N}_2$ fired at some time.

(6) $\mathcal{N}_2$ fired at every time except p.

Of these, the input described by the table of Figure 5 constitutes an occurrence of events (1), (2), (3) and (5), but not of (4), while we need to know the rest of the table to know whether it constitutes an occurrence of (6).

t	$\mathcal{N}_1$	$\mathcal{N}_2$
p	1	0
$p-1$	1	0
$p-2$	1	0

$$N_1(p)\ \&\ \overline{N_2(p)}$$
$$\&\ N_1(p-1)\ \&\ \overline{N_2(p-1)}$$
$$\&\ N_1(p-2)\ \&\ \overline{N_2(p-2)}.$$

Figure 6

5 Definite Events

5.1 "Definite Events" Defined

We shall discuss first events which refer to a fixed period of time, consisting of some ℓ (≥ 1) consecutive moments $p-\ell+1,\ldots,p$ ending with the present. We call such events *definite* of *length* (or *duration*) ℓ. Of the preceding examples, (1)–(4) are definite, but not (5) and (6).

Then in a table such as Figure 5 we need consider only the uppermost ℓ rows; e.g., that table for $\ell = 3$ then describes an event also described by the formula $N_1(p)\ \&\ \overline{N_2(p)}\ \&\ N_1(p-1)\ \&\ N_2(p-1)\ \&\ \overline{N_1(p-2)}\ \&\ N_2(p-2)$.

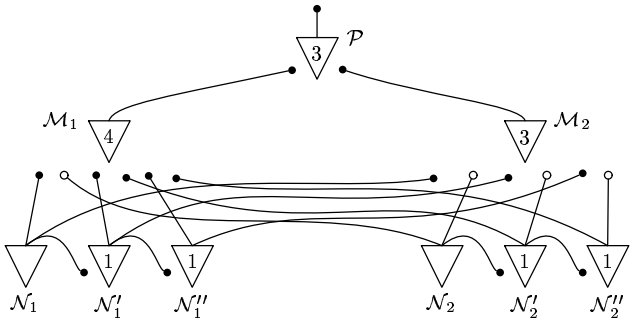
There are exactly $k\ell$ entries in a table describing the input on k neurons for the ℓ moments $p-\ell+1,\ldots,p$. Therefore there are exactly $2^{k\ell}$ possible such tables. Therefore there are exactly $2^{2^{k\ell}}$ definite events on k input neurons of length ℓ, since any particular one is determined by saying which of the inputs described by the $2^{k\ell}$ $k\times\ell$ tables would constitute an occurrence of the event.

We call a definite event *positive*, if it occurs only when at least one input neuron fires during the period to which the event refers. There are exactly $2^{2^{k\ell}-1}$ positive definite events on k input neurons of length ℓ, since that input described by the table of all 0's is excluded as an occurrence.

5.2 Representability of Definite Events: an Illustration

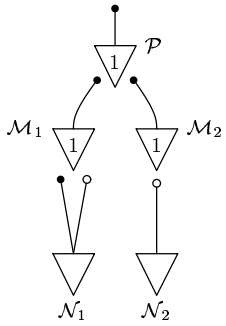
Consider the definite event which occurs when the pattern of firings fits either the table of Figure 5 (stopped at three rows) or that of Figure 6.

That is, exactly these two (out of the $2^{2\cdot 3} = 64$) 2×3 tables are to constitute an occurrence of the event. The event is described by the right member of the equivalence in Figure 7, which is obtained by combining disjunctively the conjunctions describing the respective tables separately. In the nerve net of Figure 7, the neuron $\mathcal{P}$ fires at time $p+2$, if and only if the event occurs ending at time p; or briefly, the net represents the event by the firing of $\mathcal{P}$ with *lag* 2. The neurons $\mathcal{N}_1, \mathcal{N}_1', \mathcal{N}_1''$ with just the axons connecting them are a "delay net" (cf. Figure 4). The synapse at $\mathcal{M}_1$ with the seven neurons involved is a "conjuctive net" (cf. Figure 1). That at $\mathcal{P}$



$$P(p+2) \equiv [N_1(p)\ \&\ \overline{N_2(p)}\ \&\ N_1(p-1)\ \&\ N_2(p-1)\ \&\ \overline{N_1(p-2)}\ \&\ N_2(p-2)]\ \vee\ [N_1(p)\ \&\ \overline{N_2(p)}\ \&\ N_1(p-1)\ \&\ \overline{N_2(p-1)}\ \&\ N_1(p-2)\ \&\ \overline{N_2(p-2)}].$$

Figure 7



$\mathcal{P}$ does not fire at time $p+2$, or in symbols $P(p+2) \equiv N_1(p)\ \&\ \overline{N_1(p)}$.

Figure 8

is a "disjunctive net" (cf. Figure 3).

The method of this illustration applies to every positive definite event which occurs for some one or more tables.

There remains the case of the event which never occurs. This is represented by the firing of $\mathcal{P}$ say with lag 2 in the net of Figure 8. The neuron $\mathcal{M}_2$ is inserted to show that we can have the net *connected* (in the obvious sense); otherwise $\mathcal{M}_2$ could be omitted. $\mathcal{M}_1$ could be the $\mathcal{P}$.

We have thus already proved that any positive definite event is representable by firing a neuron with lag 2. However, we shall give a more flexible treatment, listing this result as part of Theorem 1 Corollary 1.

5.3 Representability of Definite Events: General Theory

We consider logical expressions constructed using & and $\vee$ from the expressions symbolizing the firing or non-firing of one of the k input neurons $\mathcal{N}_1, \ldots, \mathcal{N}_k$ at one of the ℓ moments $p-\ell+1, \ldots, p$. Such an expression we call a *$k\ell$-formula*, of *depth* equal to the greater number of successive times & or $\vee$ is used in its construction. Here we allow any bracketing in conjunctions and disjunctions of more than two members; so e.g., $[[N_1(p)\ \&\ \overline{N_2(p)}]\ \&\ N_2(p-1)] \vee \overline{N_1(p)}$ as written is of depth 3, but it can be rewritten as $[N_1(p)\ \&\ \overline{N_2(p)}\ \&\ N_2(p-1)]\ \vee\ \overline{N_1(p)}$ with depth 2.

This definitions can be given by mathematical induction on s, thus:

1. For each i and j $(i = 1, \ldots, k; j = 1, \ldots, \ell)$, $N_i(p-j+1)$ and $\overline{N_i(p-j+1)}$ are *$k\ell$-formulas* of *depth* 0.

2. For $s > 0$, if $G_1, \ldots, G_n$ $(n \geq 2)$ are *$k\ell$-formulas* of *depth* $< s$, at least one of them being of *depth* $s-1$, then $G_1\ \&\ \ldots\ \&\ G_n$ and $G_1 \vee \ldots \vee G_n$ are *$k\ell$-formulas* of *depth* s. (Here each G_e not of depth 0 is to be enclosed in brackets when written out.)

Since the truth or falsity of a $k\ell$-formula F is determined logically from only the truth or falsity of the $N_i(p-j+1)$ which enter into it as *prime components*, each F expresses a definite event E on k input neurons of length ℓ.

We lose nothing essential by applying the negation symbol $\overline{\ }$ only directly to the prime components. For by repeated use of the logical identities $\overline{\overline{G}} \equiv G$, $\overline{G_1\ \&\ \ldots\ \&\ G_n} \equiv \overline{G_1}\ \vee\ \ldots\ \vee\ \overline{G_n}$ and $\overline{G_1 \vee \ldots \vee G_n} \equiv \overline{G_1}\ \&\ \ldots\ \&\ \overline{G_n}$, negations symbols used otherwise could be moved inward to the prime components without changing the depth. The only other operations commonly employed in the two-valued propositional calculus, namely $\rightarrow$ (*implies*) and $\equiv$, can be expressed in terms of $\overline{\ }$, & and $\vee$ thus: $G \rightarrow H \equiv \overline{G}\ \vee\ H$, $(G \equiv H) \equiv (G \rightarrow H)\ \&\ (H \rightarrow G)$.

A *circle* (of *length* c) in a nerve net is a set of distinct neurons $\mathcal{N}_1, \ldots, \mathcal{N}_c$ $(c \geq 1)$ such that $\mathcal{N}_i$ has an endbulb on $\mathcal{N}_{i+1}$ for each i $(i = 1, \ldots, c-1)$ and $\mathcal{N}_c$ has an endbulb on $\mathcal{N}_1$. The nets so far considered are without circles, including the conjuctive, disjunctive and delay nets (Figures 1, 3 and 4), and certain nets composed thence.

Theorem 1 *Let F be any $k\ell$-formula of depth s, and let E be the definite event on k input neurons of length ℓ which F expresses. There is a nerve net of structure corresponding to F (and therefore without circles) which represents E by firing or by not firing, according as E is positive or non-positive, a ceratin neuron $\mathcal{P}$ (inner if $s > 0$) at time $p+s$.*

By saying that the net is of "structure corresponding to F", we mean that it is composed out of conjunctive and disjunctive nets (together with delay nets) corresponding to the operations used in constructing F, as will be indicated in the proof.

Proof by induction on s. Under our definition of $k\ell$-formula not all of the symbols N_i $(i = 1, \ldots, k)$ need occur in F. In showing by induction how to construct the net to correspond to the logical structure of F, we incorporate only the neurons $\mathcal{N}_i$ for which N_i occurs in F. The others can be considered as floating around, unless one wishes them connected to the rest of the net, in which case if $s > 1$ they can be, e.g., as illustrated for $\mathcal{N}_2$ in Figure 8.

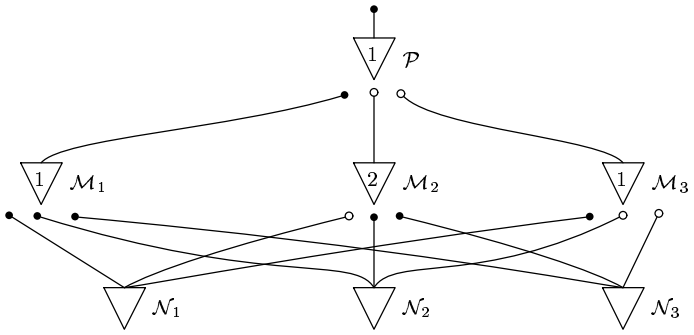
Basis: $s = 0$. Then F is $N_i(p - j + 1)$ or $\overline{N_i(p - j + 1)}$ for some i and j. Then $\mathcal{N}_i$ is also the $\mathcal{P}$, if $j = 1$; and otherwise the $\mathcal{P}$ is a neuron coming from $\mathcal{N}_i$ by a suitable delay net (Figure 4).

Induction step: $s > 0$. Then F is G_1 & ... & G_n or $G_1 \vee \ldots \vee G_n$. For $e = 1, \ldots, n$, let M_e be G_e or $\overline{G_e}$ according as (the event described by) G_e is positive or non-positive, as will be known from the case which applied to G_e. Then G_e (of depth $s_e < s$) is equivalent to M_e or $\overline{M_e}$, respectively, and by the hypothesis of the induction, there is a nerve net with a neuron $\mathcal{G}_e$ the firing or non-firing, respectively, of which at time $p + s_e$ represents G_e. Thence we obtain a neuron $\mathcal{M}_e$ the firing of which at time $p + s - 1$ represents M_e; this $\mathcal{M}_e$ is $\mathcal{G}_e$ itself if $s_e = s - 1$, and otherwise a neuron coming from $\mathcal{G}_e$ by a suitable delay net. Now we have four cases, according to how F is composed out of $M_1, \ldots, M_n$, via the construction of F from $G_1, \ldots, G_n$ and the equivalence of each G_e to one of M_e and $\overline{M_e}$.

Case 1: A conjunction containing at least one unnegated factor, e.g., M_1 & M_2 & M_3 & $\overline{M_4}$ & $\overline{M_5}$. The event is then positive; so we wish to represent it by the firing of a neuron $\mathcal{P}$ at time $p + s$. A conjunctive net (cf. Figure 1) gives us this neuron.

Case 2: A conjunction containing only negated factors, e.g., $\overline{M_1}$ & $\overline{M_2}$ & $\overline{M_3}$. The event is then non-positive. But its negation $\overline{\overline{M_1}}$ & $\overline{\overline{M_2}}$ & $\overline{\overline{M_3}}$ is positive. The latter is equivalent to $M_1 \vee M_2 \vee M_3$. A disjunctive net (cf. Figure 3) represents the latter by firing a neuron $\mathcal{P}$ at $p + s$; the net then represents the original event by the non-firing of $\mathcal{P}$ at $p + s$, which is how we wished it to be represented.

Case 3: A disjunction containing at least one negated term, e.g., $\overline{M_1} \vee \overline{M_2} \vee \overline{M_3} \vee M_4 \vee M_5$. The event is non-positive. But its negation is positive and equivalent to M_1 & M_2 & M_3 & $\overline{M_4}$ & $\overline{M_5}$. A conjunctive net represents



$$P(p+2) \equiv [N_1(p) \vee N_2(p) \vee N_3(p)] \,\&\, [N_1(p) \vee \overline{N_2(p)} \vee \overline{N_3(p)}] \,\&\, [\overline{N_1(p)} \vee N_2(p) \vee N_3(p)].$$

Figure 9

the latter by firing a neuron $\mathcal{P}$ at $p+s$; then the original event is represented by the non-firing of $\mathcal{P}$ at $p+s$.

Case 4: A disjunction containing only unnegated terms, e.g., $M_1 \vee M_2 \vee M_3$. The event is positive. A disjunctive net represents it as desired by firing a neuron $\mathcal{P}$ at $p+s$.

□

Examples The net of Figure 7 is what the present method gives for the formula there. Another illustration is in Figure 9. Treating the three formulas $N_1(p) \vee N_2(p) \vee N_3(p)$, $\mathcal{N}_1(p) \vee \overline{\mathcal{N}_2(p)} \vee \overline{\mathcal{N}_3(p)}$ and $\overline{\mathcal{N}_1(p)} \vee \mathcal{N}_2(p) \vee \mathcal{N}_3(p)$ gives us respective neurons $\mathcal{M}_1, \mathcal{M}_2$ and $\mathcal{M}_3$ which represent the events expressed, the first which is positive by firing (Case 4), the second and third which are non-positive by non-firing (Case 3), at time $p+1$. Then $\mathcal{P}$ is obtained to represent the entire event which is positive by firing at time $p+2$ (Case 1).

Corollary 1 *To each positive (non-positive) definite event, there is a nerve net without circles which represents the event by firing (not firing) a certain inner neuron at time* $p+2$.

The result was stated (for positive events and without the remark on the lag) by [McC 43].

Proof To infer this from the theorem, we need merely observe that the method of 5.2 gives a $k\ell$-formula of depth ≤ 2 to every definite event on k input neurons of length ℓ. (If the depth is < 2, a delay net may be used to increase the lag to 2.) □

Discussion Readers familiar with symbolic logic will recognize the $k\ell$-formula so obtained as a *principal disjunctive normal font* of [HIL 28]; it is "principal" because in each of its terms every one of $N_1(p), \ldots, N_k(p-\ell+1)$ occurs negated or unnegated (with an exception in the case of Figure 8).

The formula of Figure 9 is a *principal conjuctive normal form.* If the p.d.n.f. has $n < 2^{k\ell}$ terms, the p.c.n.f. has $2^{k\ell} - n$ factors. (The p.c.n.f. is obtained by evaluating the negation of the p.d.n.f. of the negation of the event.)

The use of the p.d.n.f. simplifies the proof of representability (cf. 5.2), and gives the fact that the lag can be held to 2, but the net constructed may be unnecessarily complicated. The event may admit of being described more simply by a disjunctive or conjuctive normal form not principal (which still enables the lag to be held to 2). For example (with $k = 2, \ell = 3$), $[\overline{N_2(p)}\ \&\ N_2(p-2)]\ \vee\ N_2(p-1)$ is a d.n.f., the p.d.n.f. for which would have 40 terms (the p.c.n.f. 24 factors). There may be simpler equivalents not disjunctive or conjunctive normal forms.

Since the theorem gives a representing net of corresponding structure to the formula, the problem of finding as simple nets as possible to represent definite events is correlated to the problem of finding simplest equivalents of an expression in the propositional calculus, which has recently been treated by [QUI 52].

In special cases the net may be constructed more simply than corresponding to the formula; e.g., in Figure 2 taking $p = t-1$ the net represents the event with lag 1, although the formula is of depth 3, and no equivalent formula of depth 1 exists.

Reduction of the lag below 2 is not possible in general. For example (with $k = 3, \ell = 1$) the event $N_1(p)\ \&\ (\overline{N_2(p)}\ \vee\ \overline{N_3(p)})$ is not representable with lag 1. For it is easily seen that no net consisting of a neuron $\mathcal{P}$ impinged on only by endbulbs belonging directly to $\mathcal{N}_1$, $\mathcal{N}_2$ and $\mathcal{N}_3$ can represent this event.

To hold the lag to 2, we may be obliged to have very large numbers of endbulbs originating from or impinging on a given soma.

Corollary 2 *To each positive (non-positive) definite event, there is a number s and a nerve net without circles, composed of neurons each having, and impinged upon by, at most two endbulbs and of threshold at most 2, which represents the event by the firing (non-firing) of a certain neuron at time $p+s$.*

Proof By using & and $\vee$ only as binary operations in the $k\ell$-formula, no neuron in the net construction for the theorem will be impinged upon by more than two endbulbs. Each inner neuron outside of the delay nets has only one endbulb. Each input neuron and delay net can if necessary be

replaced (increasing the lag) by a tree of neurons each with at most two endbulbs. □

We have been considering representation of an event ending with time p by the firing or by the non-firing of a certain neuron at a certain time $p+s$ ($s \geq 0$). More generally we can consider representation by a property of the state of the net (i.e., the firing or non-firing of each of its neuron) at $p+s$; i.e., the state of the net is to have or not to have this property at time $p+s$, according as the event did or did not occur ending with time p. In the following lemma and corollary, it is not being assumed that the event is definite or the net without circles.

Lemma 1 *An event which is representable in a nerve net by a property of the state at time $p+s$ for a given $s > 0$ is representable by a property of the state of the same net at time p.*

Proof What happens at times $\leq p$ can only affect the state of the net at time $p+s$ via the state of the entire net, including both the k input neurons and, say, m inner neurons, at time p. □

Corollary 3 *An event which is representable in a nerve net by a property of the state at time $p+s$ for a given $s \geq 0$ is representable by the firing or the non-firing (according to the nature of the property) of a certain inner neuron in a suitable net at time $p+2$.*

Proof We can treat the $k+m$ neurons as though all of them were input neurons, for the purpose of applying Corollary 1. By Lemma 1, the property in question is equivalent to a property of the $k+m$ neurons at time p. The latter constitutes a definite event of length 1 on the $k+m$ neurons. □

5.4 Nerve Nets without Circles

Theorem 2 *Given any nerve net without circles and given any inner neuron $\mathcal{N}$ in that net, the firing (non-firing) of that neuron at time $p+1$ is equivalent to the occurrence of a positive (non-positive) definite event.*

This theorem was stated (for positive events) by [McC 43].

Proof Whether or not $\mathcal{N}$ fires at $p+1$ is completely determined by the state (firing or non-firing) at p of those neurons $\mathcal{N}'_1, \ldots, \mathcal{N}'_r$ having endbulbs impinging on $\mathcal{N}$. Consider those of $\mathcal{N}'_1, \ldots, \mathcal{N}'_r$ which are inner neurons, and repeat the argument. Since there are no circles, any chain of neurons beginning with $\mathcal{N}$ and each impinged on by an endbulb of the next must terminate (with an input neuron). Let $\ell+1$ be the greatest of the lengths of these chains; since $\mathcal{N}$ is inner, $\ell \geq 1$. After ℓ steps, no inner neurons remain

to be considered. Thus whether or not $\mathcal{N}$ fires at time $p+1$ is completely determined by the state of certain input neurons at certain times between $p-\ell+1$ and p inclusive; i.e., $\mathcal{N}$'s firing at time $p+1$ is equivalent to a definite event of length ℓ. This event is positive, as firing can only be propagated but not originated under the law for an inner neuron's firing. □

Remark Any definite event is expressible by a logical formula, e.g., by a principal disjunctive normal form as in 5.2. So a priori there is a formula to express the event of the theorem. By utilizing the condition for firing at each synapse, which can be formulated in logical symbols depending on the threshold and the numbers and kinds of the endbulbs (cf. Figures 1–3 for several examples), one can build up a formula directly in ℓ steps, as McCulloch and Pitts indicate.

Corollary 1 *Any event which is representable in a nerve net without circles by the firing (non-firing) of a given inner neuron $\mathcal{N}$ at time $p+s$ for a given $s \geq 1$ is positive (non-positive) definite.*

Proof By the theorem, the firing of $\mathcal{N}$ at time $p+s$ is equivalent to the occurrence of a positive definite event ending with time $p+s-1$. But by the hypothesis that $\mathcal{N}$'s firing at time $p+s$ represents an event, i.e., one ending with time p (cf. Section 4), the input over the moments $p+1, \ldots, p+s-1$ has no effect on whether $\mathcal{N}$ fires at time $p+s$. □

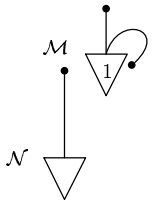
Corollary 2 *Any event which is representable in a nerve net without circles by a property of the state at time $p+s$ for a given $s \geq 0$ is definite.*

Proof By Corollary 1, with Theorem 1 Corollary 3 (which does not introduce circles). □

6 Indefinite Events: Preliminaries

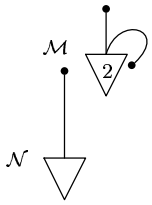
6.1 Examples

Let "(Et)" mean *there exists a t ... such that*, "(t)" mean *for all t*, and "$\rightarrow$" mean *implies*. The net in Figure 10 has a circle of length 1. If at some time $t \leq p$ the input neuron $\mathcal{N}$ fires, then $\mathcal{M}$ will fire at every subsequent moment, in particular at $p+1$ as the formula expresses. But the firing of $\mathcal{M}$ at time $p+1$ does not represent the indefinite event $(Et)_{t \leq p} N(t)$ (i.e., we do not have $M(p+1) \equiv (Et)_{t \leq p} N(t)$), if past time is infinite, because the firing of $\mathcal{M}$ at time $p+1$ can also be explained by $\mathcal{M}$ having fired at every past moment, without $\mathcal{N}$ having ever fired. Similar examples are given in Figures 11 and 12.



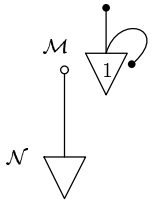
$(Et)_{t\leq p}N(t) \to M(p+1).$

Figure 10



$M(p+1) \to (t)_{t\leq p}N(t).$

Figure 11



$(Et)_{t\leq p}N(t) \to \overline{M(p+1)}.$

Figure 12

t	1	2	3	4	...
$\mathcal{N}$	0	1	1	1	...

Figure 13

This difficulty is not escapable by using other nets to represent the events, or in other examples of indefinite events, but constitutes the general rule, by Theorem 6 in Section 10 below with Lemma 1.

Of course any living organism or constructed robot has only a finite past. Theorem 6 shows that we must take this into account in the theory; otherwise we might be tempted to simplify the theory by the fiction of an infinite past, as we did in Section 5.

6.2 Initiation

Accordingly we shall hereafter assume (except when we indicate otherwise) that the past for our nerve nets goes back from p (the present) a finite time only, the first moment of which shall be 1 on our time scale. The range of the time variable in our logical formulas shall be the integers from 1 forward.

Now if in the net of Figure 10 $\mathcal{M}$ is quiet at time 1, we do have $M(p+1) \equiv (Et)_{t \leq p} N(t)$; in Figure 11 if $\mathcal{M}$ fires at time 1, $M(p+1) \equiv (t)_{t \leq p} N(t)$; and in Figure 12 if $\mathcal{M}$ fires at time 1, $\overline{M(p+1)} \equiv (Et)_{t \leq p} N(t)$. Thus the nets of Figures 10 and 12 are able to remember that $\mathcal{N}$ has fired since their beginning by changing $\mathcal{M}$ from the state it had initially; while the net of Figure 11 is able to recognize that $\mathcal{N}$ has never failed to fire by preserving $\mathcal{M}$ in the state it had originally, as [HOU 45, p. 109] have commented.

The nets in question only represent the events in question, when the inner neuron $\mathcal{M}$ has the state mentioned at time 1.

This again is the general rule for indefinite events, by Theorem 7 with Lemma 1. (Lemma 1 holds for finite past.)

We illustrate this now by showing that to represent the event $(t)_{t \leq p} N(t)$ at least one inner neuron must fire at time 1. For let the proposed representation be by a property of the state at time p (Lemma 1), i.e., solely of this state and not also of the value of p. Say $\mathcal{N}$ is the only input neuron. Were all the inner neurons quiet at $t = 1$, then with the input shown in the table of Figure 13 all inner neurons would be quiet at $t = 2$, so the state at $t = 2$ would be indistinguishable from that with Figure 14 at $t = 1$. Hence with Figure 13 the net would have the same state for $p = 2, 3, 4, \ldots$ as with Figure 14 for $p = 1, 2, 3, \ldots$, respectively, though with the former $(t)_{t \leq p} N(t)$ is false, with the latter true.

Accordingly in studying the representability of events, we shall hereafter

t	1	2	3	4	...
$\mathcal{N}$	1	1	1	1	...

Figure 14

not only choose a net for the purpose but also choose the state (firing or non-firing) of each inner neuron at time 1.

As the example of $(t)_{t\leq p}N(t)$ shows, for some events it will not be sufficient to have all the inner neurons quiet initially.

We are developing the theory of McCulloch-Pitts nerve nets as an illustrative case of the theory of finite automata. From the standpoint of the latter theory, one initial state is as reasonable as another. The alternative of excluding such events as $(t)_{t\leq p}N(p)$ from the class of representable events would be more awkward.

To one who feels that the firing of inner neurons at time 1 requires explanation under the McCulloch-Pitts laws of neural behavior, we need merely say that we have isolated certain input neurons $\mathcal{N}_1, \ldots, \mathcal{N}_k$ and a certain portion $t = 1, 2, 3, \ldots$ of time for the input for the events to be represented. We can go outside those neurons and that part of time to bring about any assumed state of our inner neurons at time $t = 1$. This is most simply accomplished by adding an extra input neuron $\mathcal{N}_{k+1}$, which is to fire at $t = 0$ and then only, and which is to have on each of the inner neurons we wish fired at $t = 1$ a number of excitatory endbulbs equal to the threshold of that neuron, but no other endbulbs.

A more complicated device, which requires an extra input neuron $\mathcal{K}$ but no extra moment of time, is illustrated in Figure 15 for the event $(t)_{t\leq p}N(p)$. The event is represented by firing $\mathcal{P}$ at $t = p + 2$ if $\mathcal{K}$ is fired, but all inner neurons are quiet, at $t = 1$. We can imagine $\mathcal{K}$ exposed to continual environmental stimulation which guarantees its firing at $t = 1$; but its firing at later times does not interfere with the representation. If $\mathcal{K}$ did not fired at $t = 1$, but first at some later time $t = u$, $\mathcal{P}$'s firing at $p+2$ would represent

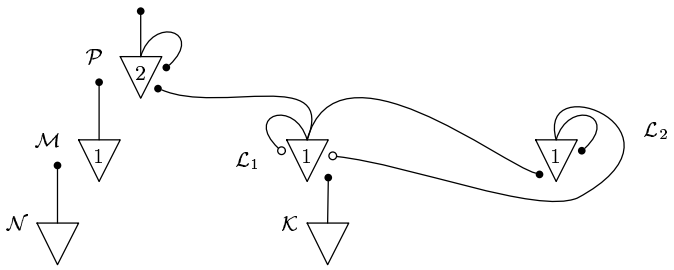


Figure 15

that $\mathcal{N}$ had fired at all moments from u to p inclusive.[1]

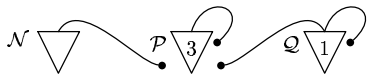
6.3 Definite Events Reconsidered

An event is a partition of the class of all the possible inputs over the past (including the present) into two subclasses, those inputs for which the event occurs and those inputs for which the event does not occur. The possible inputs on k neurons $\mathcal{N}_1, \ldots, \mathcal{N}_k$ are described by $k \times p$ tables of 0's and 1's with columns for $\mathcal{N}_1, \ldots, \mathcal{N}_k$ and rows for $t = p, \ldots, 1$. As p varies over all positive integers, these are all tables of 0's and 1's with k columns and any finite number of rows.

In Section 5 we used $k \times \ell$ tables to describe inputs over the last ℓ moments ending with the present.

Now that our time has an initial moment 1, we must be careful whenever we give a table with k columns and a finite number, say ℓ, of rows to describe an input, to make it clear whether we intend it to describe the input over the complete past (so $p = \ell$) or only over the last ℓ moments of the past (so $p \geq \ell$). In the one case we call the table *initial*, in the other *non-initial*.

[1] McCulloch and Pitts consider the problem of "solving" nets with their initial state unspecified. To "solve" for a given inner neuron $\mathcal{P}$, say at time $p+1$, means then to find for which inputs over time $1, \ldots, p$ and initial states of the inner neurons $\mathcal{P}$ will fire at time $p+1$. In the following net, the necessary and sufficient condition that $\mathcal{P}$ fire at $p+1$ is that $\mathcal{N}$ fire at all times $\leq p$ and both $\mathcal{P}$ and $\mathcal{Q}$ fire



$P(p+1) \equiv (t)_{t \leq p} N(t)\ \&\ P(1)\ \&\ Q(1).$

at time 1. This seems to be a counterexample to the formula next after (9) on [McC 43, p. 126], the proof of which we did not follow; for if we understand the formula correctly, it implies that the condition for firing of $\mathcal{P}$ should only require the existence of one neuron that fires initially. (Their 0 seems to be our 1.) This apparent counterexample discouraged us from further attempts to decipher Part III of [McC 43].

The table may be thought of as carrying a tag saying, respectively, $p = \ell$ or $p \geq \ell$. There was no necessity for this in Sections 4 and 5, as there tables referring to the complete past were infinite.

A definite event of length ℓ in Section 5 was one in which the partition of the inputs over the complete past is such that any two inputs which agree in the upper ℓ rows of their tables always fall into the same one of the subclasses. But now when $p < \ell$ there won't be ℓ rows in the table describing the input. For such a p, can the event occur? The convention we adopt is that the event shall not occur in this case. Thus the inputs of the first subclass for a definite event of length ℓ are those described by a set of non-initial $k \times \ell$ tables. If E_1 is the logical formula we used in Section 5 to describe a definite event, the event is now described by E_1 & $p \geq \ell$. The negation of this is $\overline{E_1} \vee p < \ell$, while the formula for the *complementary* definite event of length ℓ is $\overline{E_1}$ & $p \geq \ell$, which is not equivalent, except for $\ell = 1$ when the "& $p \geq \ell$" and "$\vee p < \ell$" are superfluous.

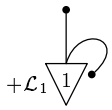
The *identical event* (written I_k or briefly I) which occurs no matter what the input (the second subclass of the partition being empty) is a definite event of length 1; the *improper event* (written $\overline{I}_k$ or $\overline{I}$) which never occurs (the first subclass being empty) can be considered as a definite event of length ℓ for every ℓ.

With the sole exception of the improper event, a given event can be definite of length ℓ for only one ℓ, and the set of the $k \times \ell$ tables (all of them non-initial) which describes it is unique. This was not the case in Section 5, as there a definite event of length ℓ was also definite of length m for each $m \geq \ell$; but now this would be absurd (except for the improper event) as the extra specification that $p \geq m$ would contradict that the event can occur for $p = \ell, \ldots, m-1$.

The definite events we have just finished describing ($2^{2^{k\ell}}$ of them for a given k and ℓ) are those which arise most naturally from those considered in Section 5 by taking into account that now the past may not include ℓ moments.

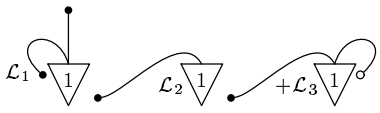
We now find it advantageous to introduce also a new kind of definite event on k neurons of length ℓ, by changing the specification for all the tables that $p \geq \ell$ to $p = \ell$; we do not include the improper event among these. These definite events we call *initial*. For a given k and ℓ, there are $2^{2^{k\ell}} - 1$ of them. An event can be an initial definite event for only one ℓ, and the set of the $k \times \ell$ tables (all of them initial) which describes it is unique. If E_1 & $p \geq \ell$ is a given non-initial definite event not improper, E_1 & $p = \ell$ is the corresponding initial definite event.

In Section 5 p entered the formulas for events only relatively, but now the events can refer to the value of p. This may seem somewhat unnatural; but, reversing the standpoint from which we were led to this in 6.1 and 6.2, if we



$L_1(p) \equiv p \geq 1.$

Figure 16



$L_1(p) \equiv p \geq 3.$

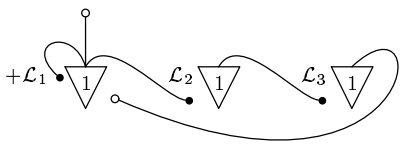
Figure 17

are to analyze nerve nets in general, starting from arbitrary initial states, we are forced to give p an absolute status. This is illustrated in Figures 16–21, where the formula gives for each net the "solution" for L_1, i.e., the condition for its firing. The "+" indicates initial firing of the indicated neuron; inner neurons not bearing a "+" are initially quiet.

Our theory can now include the case $k = 0$. (In Sections 4 and 5 we were assuming $k \geq 1$, which of course is required for nets without circles; Lemma 1 and Theorem 1 Corollary 3 hold for $k = 0$.) For $k = 0$ there are exactly three definite events of a given length ℓ, namely $p \geq \ell$, $p = \ell$ and $p \neq p$; only $p \neq p$ is positive. The nets of Figures 16–21 can be considered as representing events for $k = 0$.

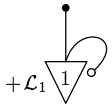
6.4 Representability of Definite Events

In Section 5 we showed how to construct nets which represent definite events on $k > 0$ input neurons of length ℓ under the assumption of an infinite past.



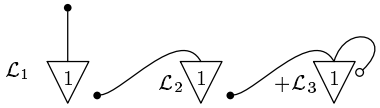
$L_1(p) \equiv p \leq 3.$

Figure 18



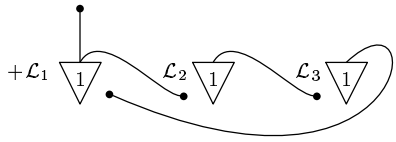
$L_1(p) \equiv p = 1.$

Figure 19



$L_1(p) \equiv p = 3.$

Figure 20



$L_1(p) \equiv [p \equiv 1 (\mathrm{mod}\ 3)].$

Figure 21

The proof there that the nets represent the events is valid now for non-initial (initial) definite events for values of $p \geq \ell$ ($p = \ell$), when the nets are started with all inner neurons quiet at time $t = 1$. To make the following discussion general, we can take the representation to be by a property of the state of the net at $t = p$ (cf. Lemma 1).

Using these nets now can sometimes give raise to a "hallucination" in the sense that the state of the net at $t = p$ has the property without the event having occurred. By reasoning similar to that in 6.2 in connection with Figures 13 and 14, this will happen for suitable inputs, when $p > \ell$ in the case of an initial definite event, and when $p < \ell$ (so $\ell > 1$) in the case of a definite event which can occur without the firing of an input neuron at its first moment $t = p - \ell + 1$.

Conversely these are the only cases in which it will happen. For consider any nerve net a property of which at $t = p$ represents a non-initial (initial) definite event correctly for $p \geq \ell$ ($p = \ell$), when the net is started at $t = 1$ with all inner neurons quiet. For there to be a hallucination when $p = m < \ell$ (so $\ell > 1$) means that for some input $\mathrm{c}_1 \ldots \mathrm{c}_m$ over $t = 1, \ldots, m$ the net has at $t = m$ a state having the property which goes with occurrence of the event. Now let the input $\mathrm{c}_1 \ldots \mathrm{c}_m$ be assigned instead to $t = \ell - m + 1, \ldots, \ell$ and an input consisting of only non-firings $\mathrm{c}'_1 \ldots \mathrm{c}'_{\ell-m}$ to $t = 1, \ldots, \ell - m$. With the input $\mathrm{c}'_1 \ldots \mathrm{c}'_{\ell-m}$ the state of the inner neurons at $t = \ell - m + 1$ must consist of all non-firings, as it did before at $t = 1$. So with the input $\mathrm{c}'_1 \ldots \mathrm{c}'_{\ell-m}\mathrm{c}_1 \ldots \mathrm{c}_m$ the net will have at $t = \ell$ the state it had before at $t = m$, which shows that $\mathrm{c}'_1 \ldots \mathrm{c}'_{\ell-m}\mathrm{c}_1 \ldots \mathrm{c}_m$ constitutes an occurrence of the event.

Call a definite event of length ℓ *prepositive*, if the event is not initial, and either $\ell = 1$ or the event only occurs when some input neuron fires at $t = p - \ell + 1$. (For $k = 0$, then only $p \geq 1$ and $p \neq p$ are prepositive.) Prepositiveness is a necessary and suficient condition for representability in a nerve net with all the inner neurons quiet initially.

This result suggests our first method for constructing nets to represent non-prepositive definite events. Say first the event is non-initial and $\ell > 1$. We supply the $\mathcal{L}_1$ of Figure 16, and treat it as though it were a $k + 1 - st$ input neuron, required to fire at $t = p - \ell + 1$ (but otherwise no taken into account) in reconsidering the event as on the $k + 1$ neurons $\mathcal{N}_1, \ldots, \mathcal{N}_k, \mathcal{L}_1$. This event on $k + 1$ neurons is prepositive, so our former methods of net construction (Section 5) apply.

A second method is to use the net of Figure 17 or 18; e.g., if the representation is by firing an inner neuron $\mathcal{P}$ (quiet at $t = 1$) at $t = p + s$, the inhibitory endbulb of $\mathcal{L}_1$ in Figure 18 shall impinge upon $\mathcal{P}$, and the figure is for $\ell + s = 5$. If the representation is by a property of the state at $t = p$, that property shall include that $\mathcal{L}_1$ of Figure 17 fire (for $\ell = 3$) or that $\mathcal{L}_1$ of Figure 18 not fire (for $\ell = 4$),

For initial definite events, the respective methods apply using Figures 19 and 20 instead of Figures 16 and 17, respectively.

The upshot is that only by reference to artifically produced firing of inner neurons at $t = 1$ could an organism recognize complete absence of stimulation of a given duration, not preceded by stimulation; otherwise it would not know whether the stimulation had been absent, or whether it had itself meanwhile come into existence.

As already remarked in 6.2, instead of an initially fired inner neuron as in Figures 16–20, we could use an additional input neuron $\mathcal{K}$ subject to continual environmental stimulation.

A hallucination of the sort considered would be unlikely to have a serious long-term or delayed effect on behaviour; but when definite events are used is building indefinite ones, this cannot be ruled out without entering into the further problem of how the representation of events is translated into overt responses.

For organisms, the picture of the nervous system as coming into total activity at a fixed moment $t = 1$ is implausible in any case. But this only means that organisms (at least those which survive) do solve their problems for their processes of coming into activity. For artificial automata or machines generally it is familiar that starting phenomena must be taken into account.

Of course our analysis need not apply to the whole experience and the entire nerve net of an organism, but $t = 1$ can be the first moment of a limited part of its experience, and the nerve net considered a sub-net of its whole nerve net.

7 Regular Events

7.1 Regular sets of tables and regular events

In this section as in 6.3 we shall use $k \times \ell$ tables (for fixed k and various ℓ, with each table tagged as either non-initial or initial) to describe inputs on k neurons $\mathcal{N}_1, \ldots, \mathcal{N}_k$ over the time $t = p - \ell + 1, \ldots, p$ for which an event shall occur. But we shall not confine our attention to the case of 6.3 that the set of tables describing when the event occurs are all of them $k \times \ell$ tables for the same ℓ and either all non-initial or all initial.

First we define three operations on sets of tables. If E and F are sets of tables, $E \vee F$ (their *sum* or *disjunction*) shall be the set of tables to which a table belongs exactly if it belongs to E or belongs to F.

If E and F are sets of tables, EF (their *product*) shall be the set of tables to which a table belongs exactly if it is the result of writing any table of F next below any non-initial table of E; if the table of E has ℓ_1 rows and that

of F has ℓ_2 rows, the resulting table has $\ell_1 + \ell_2$ rows, is initial exactly if the table of F is initial, and describes an occurrence of an event consisting in the event described by F having occurred ending with $t = p - \ell_1$, as evidenced by the input over $t = p - \ell_1 - \ell_2 + 1, \ldots, p - \ell_1$, followed by the event E having occurred ending with $t = p$, as evidenced by the input over $t = p - \ell_1 + 1, \ldots, p$. The notation EF is written so that we proceed backward into the past in reading from left to right.

Obviously $E \vee F$ and EF are associative operations. We may write E^0F for F, E^1 for E, E^2 for EE, E^3 for EEE, etc.

If E and F are sets of tables, $E * F$ (the *iterate* of E on F, or briefly E *iterate* F) shall be the infinite sum of the sets $F, EF, EEF, \ldots$, or in self-explanatory symbolism $F \vee EF \vee EEF \vee \ldots$ or $\sum_{n=0}^{\infty} E^n F$.

The *regular sets* (*of tables*) shall be the least class of sets of tables which includes the unit sets (i.e., the sets containing one table each) and the empty set and which is closed under the operations of passing from E and F to $E \vee F$, to EF and to $E * F$.

And event shall be *regular*, if there is a regular set of tables which describes it in the sense that the event occurs or not according as the input is described by one of the tables of the set or by none of them.

To include the case $k = 0$ under these definitions, we shall understand that for $k = 0$ and each $\ell \geq 1$ there are two $k \times \ell$ tables, one non-initial and one initial.[2]

Any finite set of tables is obviously regular, in particular the empty set, and the sets of $k \times \ell$ tables all with a given ℓ and either all non-initial or all initial; so every definite event is regular.

In writing expressions for regular sets or the events they describe we may omit parentheses under three associative laws ((3) – (5) in 7.2), besides which we also omit parentheses under the conventions of algebra treating $E \vee F$, EF and $E * F$ as analogous to $e + f$, ef and $e^n f$. For example, $N \vee NI * I$ means $N \vee (N(I * I))$. We may use the same letter to designate a definite event or the set of tables for it or the table itself in the case of a unit set of tables.

We write $E = F$ to say that E and F are the same set of tables; $E \equiv F$ (E is *equivalent* to F) to say that they describe the same event. Obviously $E = F \rightarrow E \equiv F$. The converse is not true in general, as we illustrate now for regular sets of tables.

Thus with $k = 1$, if N is the non-initial 1×1 table consisting of 1 (which describes the definite event symbolized in Section 5 by $N(p)$), and I is the set of all 1×1 non-initial tables (cf. 6.3), then $N \vee NI * I$ is the set of

[2] [McC 43] use a term "prehensible", introduced quite differently, but in what seems to be a related role. Since we do not understand their definition, we use another term.

non-initial $1 \times \ell$ tables (for all ℓ) with 1 in the top row. Now $N \vee NI*I \equiv N$ but $N \vee NI*I \neq N$.

We can also give counterexamples involving the distinction between non-initial and initial tables. If E is a set of tables, by E° we denote the set of tables resulting from the tables of E by redesignating as initial those which are not already initial. For any E, $E \equiv E \vee E^\circ$. But e.g., $I \neq I \vee I^\circ$.

¿From $E = F$ we can infer $E \vee G = F \vee G, G \vee E = G \vee F, EG = FG, GE = GF, E*G = F*G$ and $G*E = G*F$ by the general replacement theory of equality, since $E \vee G$, EG and $E*G$ are defined as univalent operations on sets of tables. If $=$ is replaced by $\equiv$, the third and fifth of these inferences fail to be valid in general, because the lengths of the tables in E apart from the event described by them enter into the meaning of EG and $E*G$. Thus e.g., $N \vee NI*I \equiv N$ but $(N \vee NI*I)N \neq NN$ and $(N \vee NI*I)*N \neq N*N$.

We have now two systems of notation for describing events:

(A) The logical symbolism as used for definite events is Section 5 (supplemented is Section 6 by suffixing $p \geq \ell$ or $p = \ell$) and for some examples of indefinite events in Section 6,

(B) the symbolism, usually starting with capital letters to represent definite events, by which we describe regular events (via regular sets of tables) in this section.

The question of translatability between the two systems has not yet been thoroughly investigated. By Theorem 8 in Section 12, to any expression (B) there is a logical notation (A), provided sufficient mathematical symbolism is included. Of course we have given no exact delimitation of what symbolism is to be included under (A), so the problem of translatability is not precise. In any case, with very limited mathematical symbolism included under (A), a non-regular event can be expressed, as we shall see in Section 13. It is an open problem whether there is any simple characterization of regularity of events directly in terms of the symbolism (A).

Some examples of translation from (A) to (B) follow. In the examples not involving $\mathcal{N}$, k can have any fixed value ≥ 0; involving $\mathcal{N}$ only, ≥ 1; involving $\mathcal{K}$ also, ≥ 2. Sets of non-initial $k \times 1$ tables expressing that $\mathcal{N}$ fires at time p, that $\mathcal{K}$ fires at p, and that $\mathcal{K}$ and $\mathcal{N}$ both fire at p, are denoted by N, K, and L, respectively. Also I is to be all non-initial $k \times 1$ tables; and to any set E of $k \times 1$ non-initial tables, $\overline{E}$ is the complementary set of $k \times 1$ non-initial tables, in particular $\overline{I}$ is the empty set (cf. 6.3).

$(Et)_{t\le p}N(t)$	$I * N$
$(t)_{t\le p}N(t)$	$N * N^\circ$
$(Eu)_{u\le p}[K(u)\ \&\ (t)_{u\le t\le p}N(t)]$	$N * L$
$(Eu)_{u\le p}[K(u)\ \&\ (s)_{s<u}\overline{K(s)}\ \&\ (t)_{u\le t\le p}N(t)]$	$N * L^\circ\ \vee\ N * L * \overline{K} * \overline{K}^\circ$
$N(t)$ for at least two values of $t \le p$	$I * NI * N$
$N(t)$ for exactly one value of $t \le p$	$\overline{N} * N^\circ\ \vee\ \overline{N} * N\overline{N} * \overline{N}^\circ$, call this M
$N(t)$ for an odd number of values of $t \le p$	$(\overline{N} * N\overline{N} * N) * M$
$p \ge 3$	I^3
$p = 1$	I°
$p \equiv 1 (\mathrm{mod}\ \ 3)$	$(I^3) * I^\circ$
$p \le 3$	$I^\circ\ \vee\ II^\circ\ \vee\ I^2I^\circ$

7.2 Algebraic Transformations of Regular Expressions

We list some equalities for sets of tables. (We have scarcely begun the investigation of equivalences.)

(1) $E\ \vee\ E = E$.

(2) $E\ \vee\ F = F\ \vee\ E$.

(3) $(E\ \vee\ F)\ \vee\ G = E\ \vee\ (F\ \vee\ G)$.

(4) $(EF)G = E(FG)$.

(5) $(E * F)G = E * (FG)$.

(6) $(E\ \vee\ F)G = EG\ \vee\ FG$.

(7) $E(F\ \vee\ G) = EF\ \vee\ EG$.

(8) $E * (F\ \vee\ G) = E * F\ \vee\ E * G$.

(9) $E * F = F\ \vee\ E * EF$.

(10) $E * F = F\ \vee\ EE * F$.

(11) $E * F = E^s * (F\ \vee\ EF\ \vee\ E^2F\ \vee\ \ldots\ \vee\ E^{s-1}F) \quad (s \ge 1)$.

(12) $E\ \vee\ \overline{I} = \overline{I}\ \vee\ E = E$.

(13) $E\overline{I} = \overline{I}E = \overline{I}$.

(14) $E * \overline{I} = \overline{I}$.

(15) $\overline{I} * E = E$.

(16) $E^\circ \vee F^\circ = (E \vee F)^\circ$.

(17) $EF^\circ = (EF)^\circ$.

(18) $E^\circ F = \overline{I}$.

(19) $E * F^\circ = (E * F)^\circ$.

(20) $E^\circ * F = F$.

(21) $(E \vee F^\circ) * G = E * G$.

(22) $E^{\circ\circ} = E^\circ$.

(23) $\overline{I}^\circ = \overline{I}$.

To prove (11), we have

$$E * F = \sum_{n=0}^{\infty} E^n F = \sum_{q=0}^{\infty} \sum_{r=0}^{s-1} E^{sq+r} F = \sum_{q=0}^{\infty} E^{sq} \sum_{r=0}^{s-1} E^r F.$$

In this subsection, we shall deal with particular ways of expressing a regular set of tables under the definition in 7.1. As we saw there, we can as well start with any sets of tables for definite events, instead of simply with the unit sets and the empty set. By a *regular expression*, we shall mean a particular way of expressing a regular set of tables starting with sets of tables for definite events and applying zero or more times the three operations (passing from E and F to $E \vee F$, EF or $E * F$); the occurrences of sets of tables for definite events with which the construction starts we call them *units*. A unit is of *length* ℓ, if the definite event described by it is of length ℓ; *initial*, if that is *initial*. (It comes to almost the same to let "regular expression" mean a notation for a regular set of tables obtained by starting with symbols for definite events and combining them by use of the notations "$E \vee F$", "EF" and "$E * F$", and most of what we say can be read either way. But when we say that $\overline{I}$ does not occur in a regular expression as a unit, in terms of notation we would mean that neither "$\overline{I}$" nor any other symbol for the empty set occurs as a unit. Also, in terms of notation we would have to identify the units whenever they are not all of them single letters.)

Lemma 2 *Each regular expression is reducible either to* $\overline{I}$ *or to a regular expression in which* $\overline{I}$ *does not occur as a unit.*

Proof By repeated use of (12) – (15). □

Lemma 3 *Each regular expression* G *is reducible to the form* $G_1 \vee G_2^\circ$ *where* G_1 *contains no initial units.*

Proof, by induction on the number n of units in G.

Basis: $n = 1$. Then G is a unit. If G is not initial, let $G_1 = G$ and use (12) and (23). If G is initial, let $G_2^\circ = G$ and use (12).

Induction step: $n > 1$.

Case 1: G is $E \vee F$. By the hypothesis of the induction, $E = E_1 \vee E_2^\circ$ and $F = F_1 \vee F_2^\circ$. Thence, using (2), (3) and (16), $G = (E_1 \vee F_1) \vee (E_2 \vee F_2)^\circ$.

Case 2: G is EF. Using the hypothesis of the induction, (6) and (7), (18) and (12), and (17), $G = E_1F_1 \vee (E_1F_2)^\circ$.

Case 3: G is $E * F$. By the hypothesis of the induction, (21), (8) and (19), $G = E_1 * F_1 \vee (E_1 * F_2)^\circ$.

□

We define recursively the "earliest units" of a regular expression, thus.

1. A regular expression consisting of only one unit is its *earliest unit.*
2. The *earliest units* of E and the *earliest units* of F are the *earliest units* of $E \vee F$.
3. The *earliest units* of F are the earliest units of EF and of $E * F$.

Lemma 4 *Each regular expression G is reducible to $\overline{I}$ or to a regular expression in which $\overline{I}$ does not occur as a unit and only earliest units are initial.*

Proof, by induction on the number n of units in G.

Basis: $n = 1$. Then G is $\overline{I}$, or not $\overline{I}$ but a unit and therefore earliest.

Induction step: $n > 1$.

Case 1: G is $E \vee F$. Use the hypothesis of the induction.

Case 2: G is EF. By Lemma 3, $E = E_1 \vee E_2^\circ$. Thence, using (6), (18) and (12), $G = E_1F$. Now apply the hypothesis of the induction to F.

Case 3: G is $E * F$. Using Lemma 3 and (21), $G = E_1 * F$. Apply the hypothesis of the induction to F.

□

In transforming regular expressions we may reconstitute the units; e.g., when E_1 and E_2 are units of lengths ℓ_1 and ℓ_2, E_1 not being initial and neither being $\overline{I}$, then E_1E_2 can be taken as a new unit, which is of length $\ell_1 + \ell_2$ and initial or not according as E_2 is initial or not.

Lemma 5 *For each $s \geq 1$: Each regular expression is reducible either to $\overline{I}$ or else to a regular expression not containing $\overline{I}$ as a unit, having initial units only as earliest units, and having the form of a disjunction of one or more terms of two kinds: a unit of length $< s$, or a regular expression composed of units all of length $\geq s$.*

One can always take the number of terms of the second kind to be one, since a disjunction of terms of the second kind is a term of the second kind (cf. (2) and (3)).

Proof For $s = 1$, Lemma 5 coincides with Lemma 4. Now take a fixed $s \geq 2$, and suppose that after applying Lemma 4 we have a regular expression G of the second type there. Transformation of G by any of (1)–(11), which include all individual transformation steps used in what follows, preserves this type. This enables us below to reconstitute $E_1E_2 \ldots E_m$, where $E_1, \ldots, E_m$ are units, as a new unit. Now we show by induction on the number n of units in G, that G can be transformed into a disjunction of terms of the two kinds for Lemma 5.

Basis: $n = 1$. Then G is of the first or second kind according as its length is $< s$ or $\geq s$.

Induction step: $n > 1$.

Case 1: G is $E \vee F$. Then E and F will each be of the second type of Lemma 4. So by the hypothesis of the induction, E and F are both expressible as disjunctions of terms of only the two kinds. Thence so is $E \vee F$, by combining the two disjunctions as one disjunction.

Case 2: G is EF. Using the hypothesis of the induction, (6) and (7), EF is then equal to a disjunction of terms each of which is of one of the four types

$$E'F', \; E''F'', \; E''F', \; E'F''$$

where $'$ identifies a factor (originally a term of the disjunction for E or for F) of the first kind, $''$ of the second kind. By the reasoning of Case 1, it will suffice to show that each of these four types of products is expressible as a disjunction of terms of the two kinds.

But $E'F'$ can be reconstituted as a unit, and according as this new unit is of length $< s$ or $\geq s$, $E'F'$ becomes of the first or of the second kind.

The product $E''F''$ is of the second kind.

Now consider $E''F'$. Using (4)–(6), the F' can be moved progressively inward until finally F' occurs only in parts of the form HF' where H is a unit of length $\geq s$. Each such part can be reconstituted as a unit of length $\geq s+1$, so that $E''F'$ becomes of the second kind.

For $E'F''$ we proceed similarly, using (4) (from right to left), (7), and (10) followed by (7) and (4).

Case 3: G is $E * F$. Applying to $E * F$ successively (11) and (9), $E*F = F \vee EF \vee E^2F \vee \ldots \vee E^{s-1}F \vee E^s * E^s(F \vee EF \vee E^2F \vee \ldots \vee E^{s-1}F)$. Since $E^2F, \ldots, E^{s-1}F$ are simply repeated products, by the method of Case 2 (repeated as necessary) each of $F, EF, E^2F, \ldots, E^{s-1}F$ is expressible as a disjunction of terms of the two kinds. Now consider E^s; by taking the disjunction for E given by the hypothesis of the induction, and multiplying out ((6) and (7)), we obtain a sum of products of s factors each as in Case 2 we obtained a sum of products of 2 factors each. A product in which the factors are all of the first kind can be reconstituted as a unit, which will be of length $\geq s$ since the number of the factors is s, so it becomes of the second kind. All other types of products which can occur include a factor of the second kind, so by the treatment of the three types of products $E''F''$, $E''F'$ and $E'F''$ under Case 2 (repeated as necessary), each of these becomes of the second kind. So E^s (after suitable reconstitution of units) becomes of the second kind. Now by the treatment of the two types of products $E''F''$ and $E''F'$ under Case 2, $E^s(F \vee EF \vee E^2F \vee \ldots \vee E^{s-1}F)$ and hence $E^s * E^s(F \vee EF \vee E^2F \vee \ldots \vee E^{s-1}F)$ become of the second kind.

□

Lemma 6 *Each regular expression is reducible without reconstituting the units to a disjunction of one or more terms of the form E_iF_i where each E_i is a unit and F_i is empty (then E_iF_i is E_i) or regular (then E_i is non-initial).*

Proof For a regular expression G of the second type of Lemma 4, one can see, by induction on the number n of units in G, that G can be transformed (using only (4), (6), (10)) into a disjunction of terms of the two kinds for Lemma 6. □

7.3 Representability of regular events

A $k \times \ell$ table is *prepositive* (*positive*), if it describes a prepositive definite event 6.4, i.e., if it is not initial and either $\ell = 1$ or there is a 1 in its lowest row (a positive definite event 5.1, 6.3, i.e., if there is a 1 is some row). A set of tables is *prepositive* (*positive*), if every table of the set is prepositive (positive).

Theorem 3 *To each regular event, there is a nerve net which represents the event by firing a certain inner neuron at time $p + 2$, when started with suitable states of the inner neurons at time 1. If the event is describable by a prepositive and positive regular set of tables, the representation can be by a net started with all inner neurons quiet.*

Proof We begin by establishing the theorem for a regular event described by a term G of the second kind for Lemma 5 with $s = 2$. We use induction on the number n of units in G.

We will arrange that the neuron (call it the *output neuron*) which is to fire at $p + 2$ exactly if the event occurs ending with time p shall be of threshold 1 impinged on by only excitatory endbulbs (as in Figure 3), and shall have no axons feeding back into the net.

Basis: $n = 1$. We construct a net to represent (the event described by) G by the method of proof of Theorem 1 Corollary 1 if G is prepositive (a fortiori positive, since $\ell \geq s > 1$), and otherwise this with the first method of 6.4 (with a neuron of Figure 16 or Figure 19) which makes the event prepositive (so positive) as an event on $k+1$ neurons, so the representation is by firing at time $p + 2$ in both cases.

Induction step: $n > 1$.

Case 1: G is $E \vee F$. By the hypothesis of the induction, there are nets which represent E and F, respectively, each in the manner described, say with respective output neurons P and Q. To represent $E \vee F$, we "identify" P and Q, i.e., we replace them by a single neuron (call it P) having all the endbulbs which impinged separately on P and on Q; and of course we similarly identify the input neurons $N_1, \ldots, N_k$. The resulting net is diagramed in Figure 22. The box marked $\mathcal{E}$ stands for the net for E except for its input neurons and output neuron. The heavy line leading to P from this box represents the axons which formerly led to the output neuron $\mathcal{P}$ in the net for E.

Case 2: G is EF. Consider the expression E' which is obtained from E by altering each unit to make it refer to one new input neuron

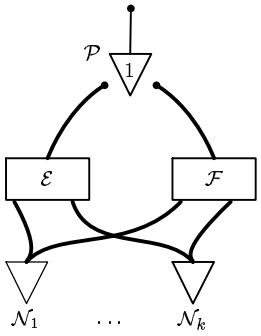


Figure 22 $E \vee F$

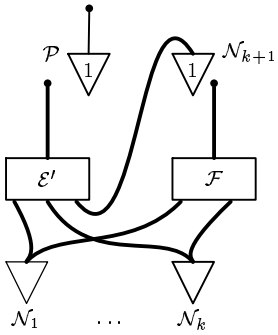


Figure 23 EF

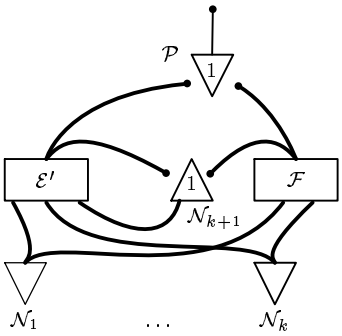


Figure 24 $E * F$

$\mathcal{N}_{k+1}$ required to fire at the second moment of each earliest unit (but otherwise not affecting the occurrence or non-occurrence of the respective definite events); by the hypothesis that the units are of length ≥ 2, there is a second moment. Then E' is of the second kind for Lemma 5 with the same number of units as E, since the alteration gives a regular expression with the same structure in terms of its respective units under the three operations. So by the hypothesis of the induction we can represent E' and F by nets as described. However we simplify the construction by leaving out the neuron of Figure 16 in the case of each earliest non-prepositive unit of E' (this unit is non-initial, by one of the properties of G secured in Lemma 5). Now the net for EF is obtained by identifying the new input neuron $\mathcal{N}_{k+1}$ in the net for E' with the output neuron $\mathcal{Q}$ of the net for F, besides of course identifying the input neurons $\mathcal{N}_1, \ldots, \mathcal{N}_k$ for the two nets, and taking as output neuron the output neuron $\mathcal{P}$ for E' (Figure 23). The event E' is positive, since $\mathcal{N}_{k+1}$ is required to fire at its second moment. No hallucination is possible as a result of leaving out the neurons of Figure 16 for earliest non-prepositive units of E', since $\mathcal{N}_{k+1}$ (required for those units to fire at their second moment) cannot fire until two moments after an occurrence of F, by the construction of the net for F. These omissions of neurons of Figure 16 are to give the last statement of the theorem.

Case 3: G is $E * F$. The nets for E' and F are combined as in Figure 24.

□

Conclusion This completes the induction to show the representability of a regular event described by a term G of the second kind for Lemma 5 with $s = 2$. Terms of the first kind are treated as under the basis (but using Figure 16 additionally in the case with $\ell = 1$ of a prepositive non-positive term), and the disjunction of terms (if there are more than one) as under Case 1. The case the event is $\overline{I}$ has already been treated in Section 5 (Figure 8).

Discussion If the original regular expression for the event is already in terms of units each of length ≥ 2, the proof of the theorem is straightforward and yields nets of complexity corresponding very well to that of the regular expression. (For simplification of the nets representing the units, possibly at the cost of increasing the lag above 2, cf. the discussion following Theorem 1 Corollary 1.) The difficulty which calls for complicated reformulation via the proof of Lemma 5 arises when we try to combine in succession the

representations of events some of them shorter than the time necessary for the net to organize a representation of the preceding event by the firing of a single neuron; the solution by Lemma 5 consists in considering grosser events before trying to combine the representations.

Theorem 4 *To each event constructed from regular events by the operations* $\&$, $\vee$, $-$, *there is a nerve net which represents the event by firing a certain inner neuron at time* $p+2$, *when started with suitable states of the inner neurons at time* 1.

The proof will follow. By Corollary Theorem 5 below, all representable events are regular. So by Theorem 4 and 5 together, combinations of regular events by $\&$, $\vee$ and $-$ are regular, which with Theorem 3 includes Theorem 4. We have not defined $\&$ and $-$ as operations on sets of tables, so EF and $E * F$ cannot be used after the application of $\&$ or $-$.

Proof of theorem 4 To each of the regular events which enter in the construction by $\&$, $\vee$ and $-$ of the given event, consider a regular expression for the regular event. Apply to this Lemma 5 with $s = 2$, and to the resulting terms of the second kind Lemma 6. Thus we obtain an expression for the given event by the operations $\&$, $\vee$ and $-$ from components $E_1F_1, \ldots, E_mF_m$ where each E_i is an expression for a definite event and F_i is a regular expression (then the definite event expressed by E_i is non-initial and of length ≥ 2) or empty. Let E_i' come from E_i as E' come from E in the proof of Theorem 3 Case 2 if F_i is regular, and be the result of introducing an extra input neuron $\mathcal{N}_{k+i}$ to fire at the first moment of E_i' if F_i is empty. Now consider (as an event on the $k+m$ neurons $\mathcal{N}_1, \ldots, \mathcal{N}_k, \mathcal{N}_{k+1}, \ldots, \mathcal{N}_{k+m}$) the same combination of $E_1', \ldots, E_m'$ as the given event is of $E_1F_1, \ldots, E_mF_m$. If this combination of $E_1', \ldots, E_m'$ when treated as a definite event in the sense of Section 5 (not Section 6) of length equal to the greatest of the lengths of $E_1', \ldots, E_m'$ is not positive, we make it so by adding "& E_{m+1}'" where E_{m+1}' refers to the firing of a neuron $\mathcal{N}_{k+m+1}$ at time p. Now use the method of net construction for Theorem 1 Corollary 1 to construct a representing net for this event on $k+m$ or $k+m+1$ neurons. Then for each i for which F_i is regular, identify $\mathcal{N}_{k+i}$ with the output neuron of a net given by Theorem 3 representing F_i; and for each i for which F_i is empty make $\mathcal{N}_{k+i}$ an inner neuron required to fire at time 1, as in Figure 16 if E_i is non-initial or $i = m+1$, and as in Figure 19 if E_i is initial. □

7.4 Problems

Numerous problems remain open, which the limited time we have given to this subject did no permit us to consider, although probably some of them at least can be solve quickly.

Is there an extension of Theorem 1 Corollary 2 to all regular events?

By the *complete set of tables* for an event we mean the set of tables all of them initial which describes the event. By the *minimal set of tables* for an event we mean the set of tables describing the event each of which has the property that neither a proper upper segment of it, nor itself if it is initial, as a non-initial table describes an occurrence of the event. The complete set of tables for a regular event is regular, by Theorem 3 and the proof of Theorem 5. Is the minimal set necessarily regular? If so, can a regular expression for it be obtained effectively from a regular expression for the complete set?

What kinds of events described originally in other terms are regular? We have only some examples of translation from (A) to (B) (end 7.1), and one of an indirectly established closure property of regular events (Theorem 4 with Theorem 5).

Given a regular expression for an event, it may be difficult to see of what the event consists. We know cases in which a very complicated regular expression is equivalent to a much simpler one, e.g., some arising via the proof of Theorem 5. Are there simple normal forms for regular expressions, such that any regular expression is equal, or is equivalent, to one in a normal form? Is there an effective procedure for deciding whether two regular expressions are equal, or are equivalent?

Our reason for introducing the regular events, as given by regular sets of tables described by regular expressions, is Theorem 5, which we discovered before Theorem 3. By using the notion of regular events, we thus demonstrate that a McCulloch-Pitts nerve net can represent any event which any other kind of finite digital automaton (in the sense to be developed in detail in Section 8) can represent. This of course includes a number of special results which McCulloch and Pitts obtained for alternative kinds of nerve nets, but is more general. The way is open to attempt similarly to verify the like for other kinds of "cells" in place of neurons, or to seek some characterization of the properties of the cells in order that aggregates of them have the capacity for representing all representable (i.e., all regular) events.

PART II. FINITE AUTOMATA

8 The Concept of a Finite Automaton

8.1 Cells

Time shall consist of a succession of discrete moments numbered by the positive integers, except in Section 10 where all the integers will be used.

We shall consider automata constructed of a finite number of parts called *cells*, each being at each moment in one of a finite number ≥ 2 of states.

We shall distinguish two kinds of cells, *input cells* and *inner cells.*

An input cell admits two states, 0 and 1 (or "quiet" and "firing"), which one is assumed at a given moment being determined by the environment.

The restriction to 2 states for input cells makes the notion of an input to the automaton coincide with the notion of an input to a nerve net as formulated in Sections 4 and 6.3. But the present theory would work equally well with more then 2 states. Nothing would be gained, however, as p cells admitting each admitting 2 states could be used to replace one cell admitting any number q $(2 \leq q \leq 2^p)$ of states $0, 1, \ldots, q-1$, where if $q < 2^p$ we could either consider only inputs in which states $q, \ldots, 2^p - 1$ do not occur or identify those states with the state $q-1$ in all the operations of the automaton.

The number of states of an inner cell is not restricted to 2, and different inner cells may have different numbers of states.

The state of each inner cell at any time $t > 1$ is determined by the states of all the cells at time $t-1$. Of course it may happen that we do not need to know the states of all the cells at time $t-1$ to infer the state of a given inner cell at time t. Our formulation merely leaves it unspecified what kind of a law of determination we use, except to say that nothing else than the states of the cells at $t-1$ can matter.

For time beginning with 1, the state of each of the inner cells at that time is to be specified (except in Section 11).

A particular example of a finite automaton is a McCulloch-Pitts nerve net (Part I). Here all the cells admit just 2 states. Another example is obtained by considering inner neurons with "alterable endbulbs" which are not effective unless at some previous time the neuron having the endbulb and the neuron on which the endbulb impinges were simultaneously fired. A neuron with r such alterable endbulbs admits 2^{r+1} states. Many other possibilities suggest themselves.

8.2 State

With k input cells $\mathcal{N}_1, \ldots, \mathcal{N}_k$ $(k \geq 0)$, and m inner cells $\mathcal{M}_1, \ldots, \mathcal{M}_m$ $(m \geq 1)$ with respective numbers of states $s_1, \ldots, s_m$, there are exactly $2^k \cdot s_1 \cdot \ldots \cdot s_m$ possible (*complete*) *states* of the automaton. We can consider each states as a combination of an *external state*, of which there are 2^k possible, and an *internal state*, of which there are $s_1 \cdot \ldots \cdot s_m$ possible.

The law by which the states of the inner cells at time $t > 1$ are determined by the states of all the cells at time $t-1$ can be given by specifying to each of the complete states at time $t-1$ which one of the internal states at time

t shall succeed it.

We could indeed consider the entire aggregate of m internal cells as replaced by a single one admitting $s_1 \cdot \ldots \cdot s_m$ states. We shall not take advantage of this possibility, because we have in view applications of the theory of finite automata in which the cells have certain simple properties and are connected in certain simple ways.

We could also (but shall not) get along with a single input cell, by scheduling the inputs on the k original input cells to come in successively in some order on the new one, which would alter the time scale so that k moments of the new scale correspond to 1 of the original. Events referring to the new time scale could then be interpreted in terms of the original.

Now let us call the states $\mathfrak{a}_1, \ldots, \mathfrak{a}_r$ where $r = 2^k \cdot s_1 \cdot \ldots \cdot s_m$ and the internal states $\mathfrak{b}_1, \ldots, \mathfrak{b}_q$ where $q = s_1 \cdot \ldots \cdot s_m$. Let the notation be arranged so that the internal state at time 1 is $\mathfrak{b}_1$.

With the internal state at time 1 fixed, the state at time p is a function of the input over the time $1, \ldots, p$ (including the value of p, or when $k = 0$ only this).

So each of the states $\mathfrak{a}_1, \ldots, \mathfrak{a}_r$ represents an event, which occurs ending with time p, if and only if the input over the time $1, \ldots, p$ is one which results in that one of $\mathfrak{a}_1, \ldots, \mathfrak{a}_r$ being the state at time p. Thus the automaton can know about its past experience (inclusive of the present) only that it falls into one of r mutually exclusive classes (possibly some of them empty).

Similarly an internal state at time $p + 1$, or a property of the complete state at time p, or a property of the internal state at time $p+1$, or a property of the internal state at time $p + s$ for an $s > 1$ which does not depend on the input over the time $p+1, \ldots, p+s-1$, represents an event. Thus to say that the state at time p has a certain property is to say that the state then is one of a certain subclass of the r possible states, so that the past experience falls into the set sum (or disjunction) of the representative classes of past experiences which are separately represented by the states of the subclass.

What sorts of events can be represented? As the concept of input is the same as in Part I, we can use the notion of "regular event" which was introduced in Section 7. The following theorem, together with Theorem 3 referring to a special kind of finite automaton, answer the question.

9 Regularity of Representable Events

Theorem 5 *In any finite automaton (in particular, in a McCulloch-Pitts nerve net), started at time* 1 *in a given internal state* $\mathfrak{b}_1$*, the event represented by a given state existing at time* p *is regular.*

Proof Since the initial internal state is specified, there are 2^k possible

initial states (the results of combining the given initial internal state $\mathfrak{b}_1$ with each of the 2^k possible external states at time 1).

So if we can show that the automaton starting from a given state at time 1 will reach a given state at time p, if and only if a certain regular event occurs ending with time p, then the theorem will follow by taking the disjunction of 2^k respective regular events, which is itself a regular event. □

Given any state $\mathfrak{a}$ at time $t-1$ ($t \geq 2$), exactly 2^k states are possible at time t, since the internal part of the state at time t is determined by $\mathfrak{a}$, and the external part can happen in 2^k ways. Let us say each of these 2^k states is *in relation* R to $\mathfrak{a}$.

The next part of our analysis will apply to any binary relation R defined on a given set of $r \geq 1$ objects $\mathfrak{a}_1, \ldots, \mathfrak{a}_r$ (called "states"), whether or not it arises in the manner just described.

Consider any two $\mathfrak{a}$ and $\overline{\mathfrak{a}}$ of the states, not necessarily distinct. We shall study the strings of states $\mathfrak{d}_p \mathfrak{d}_{p-1} \ldots \mathfrak{d}_1$ ($p \geq 1$) for which $\mathfrak{d}_p$ is $\mathfrak{a}$, $\mathfrak{d}_1$ is $\overline{\mathfrak{a}}$, and for each t ($t = 2, \ldots, p$) $\mathfrak{d}_t$ is in relation R to $\mathfrak{d}_{t-1}$ (in symbols, $\mathfrak{d}_t \ R \ \mathfrak{d}_{t-1}$); say such strings *connect* $\mathfrak{a}$ to $\overline{\mathfrak{a}}$.

We say a set of strings is "regular" under the following definition (chosen analogously to the definition of "regular" sets of tables in 7.1).

The empty set and for each i ($i = 1, \ldots, r$) the unit set $\{\mathfrak{a}_i\}$ having as only member $\mathfrak{a}_i$ considered as a string of length 1 are *regular*. If $\mathcal{A}$ and $\mathcal{B}$ are *regular*, so is their sum, written $\mathcal{A} \vee \mathcal{B}$. If $\mathcal{A}$ and $\mathcal{B}$ are *regular*, so is the set, written $\mathcal{AB}$, of the strings obtainable by writing a string belonging to $\mathcal{A}$ just left of a string belonging to $\mathcal{B}$. If $\mathcal{A}$ and $\mathcal{B}$ are *regular*, so is the sum, written $\mathcal{A} * \mathcal{B}$, for $n = 0, 1, 2, \ldots$ of the sets $\mathcal{A} \ldots \mathcal{AB}$ with n $\mathcal{A}$'s preceding the $\mathcal{B}$.

Lemma 7 *The strings* $\mathfrak{d}_p \ldots \mathfrak{d}_1$ *connecting* $\mathfrak{a}$ *to* $\overline{\mathfrak{a}}$ *constitute a regular set.*

Proof of Lemma, by induction on r.

Basis: $r = 1$. Then $\overline{\mathfrak{a}}$ is $\mathfrak{a}$. If $\overline{\mathfrak{a} R \mathfrak{a}}$ (i.e., if R is an irreflexive relation), the set of the strings connecting $\mathfrak{a}$ to $\mathfrak{a}$ is the unit set $\{\mathfrak{a}\}$, which is regular. If $\mathfrak{a} R \mathfrak{a}$, then the set is $\{\mathfrak{a}, \mathfrak{a}\mathfrak{a}, \mathfrak{a}\mathfrak{a}\mathfrak{a}, \ldots\}$, which is regular, since it can be written $\mathcal{A} * \mathcal{A}$ where $\mathcal{A} = \{\mathfrak{a}\}$.

Induction step: $r > 1$.

Case 1: $\mathfrak{a} = \overline{\mathfrak{a}}$. In this case any string connecting $\mathfrak{a}$ to $\overline{\mathfrak{a}}$ is of the form

$$\mathfrak{a} \to \mathfrak{a} \to \mathfrak{a} \to \ldots \mathfrak{a} \to \mathfrak{a},$$

where the number of $\mathfrak{a} \rightarrow$'s is ≥ 0 and each $\rightarrow$ represents independently the empty string (this being possible only if $\mathfrak{a}R\mathfrak{a}$) or a non-empty string without any $\mathfrak{a}$ in it. Let $\mathfrak{e}_1, \ldots, \mathfrak{e}_g$ $(g \geq 0)$ be the states $\mathfrak{e}$ such that $\mathfrak{a}R\mathfrak{e}$ but $\mathfrak{e} \neq \mathfrak{a}$, and $\mathfrak{b}_1, \ldots, \mathfrak{b}_h$ $(h \geq 0)$ the states $\mathfrak{b}$ such that $\mathfrak{b}R\mathfrak{a}$ but $\mathfrak{b} \neq \mathfrak{a}$. Now any non-empty string represented by an $\rightarrow$ must start with one of $\mathfrak{e}_1, \ldots, \mathfrak{e}_g$ and end with one of $\mathfrak{b}_1, \ldots, \mathfrak{b}_h$. For each pair $\mathfrak{e}_i\mathfrak{b}_j$, by the hypothesis of the induction the set of the strings connecting $\mathfrak{e}_i$ to $\mathfrak{b}_j$ without $\mathfrak{a}$ in it is regular. Say $\mathcal{B}_1, \ldots, \mathcal{B}_{gh}$ are these regular sets; and let $\mathcal{A}$ be $\{\mathfrak{a}\}$. Now if $\mathfrak{a}R\mathfrak{a}$, the set of the possible strings $\mathfrak{a} \rightarrow$ is $\mathcal{A} \vee \mathcal{A}(\mathcal{B}_1 \vee \ldots \vee \mathcal{B}_{gh})$ (which reduces to $\mathcal{A}$ if $gh = 0$ or all the $\mathcal{B}$'s are empty); and if $\overline{\mathfrak{a}R\mathfrak{a}}$, it is $\mathcal{A}(\mathcal{B}_1 \vee \ldots \vee \mathcal{B}_{gh})$ (which is empty if $gh = 0$ or all the $\mathcal{B}$'s are empty). Let this set be $\mathcal{C}$. Then the set of the strings leading from $\mathfrak{a}$ to $\mathfrak{a}$ is $\mathcal{C} * \mathcal{A}$ (which reduces to $\mathcal{A}$ if $\mathcal{C}$ is empty).

Case 2: $\mathfrak{a} \neq \overline{\mathfrak{a}}$. Now we have instead

$$\mathfrak{a} \rightarrow \mathfrak{a} \rightarrow \mathfrak{a} \rightarrow \ldots \mathfrak{a} \rightarrow \mathfrak{a} \rightsquigarrow \overline{\mathfrak{a}},$$

where the number of $\rightarrow$'s is ≥ 0 and each $\rightarrow$ and the $\rightsquigarrow$ represents independently the empty string or a non-empty string without any $\mathfrak{a}$ in it. If $\mathcal{D}$ is the set of the possible strings $\mathfrak{a} \rightsquigarrow$, and $\mathcal{E} = \{\overline{\mathfrak{a}}\}$, the set of the strings connecting $\mathfrak{a}$ to $\overline{\mathfrak{a}}$ is $\mathcal{C} * \mathcal{D}\mathcal{E}$, which is regular.

□

Proof of theorem (completed) We need to show that, for a given state $\mathfrak{a}$ and each of 2^k states $\overline{\mathfrak{a}}$, the state is $\mathfrak{a}$ at time p and $\overline{\mathfrak{a}}$ at time 1, if and only if a certain regular event occurs over the time $1, \ldots, p$.

By the lemma, the set of the strings which can connect $\mathfrak{a}$ to $\overline{\mathfrak{a}}$ is regular. Consider an expression for this regular set in terms of the empty set and the sets $\{\mathfrak{a}_i\}$ as the units (cf. 7.2). In this expression let us replace each unit $\{\mathfrak{a}_i\}$ by the unit set consisting of the $k \times 1$ table which (if $k > 0$) describes the external part of the state $\mathfrak{a}_i$, labeled initial or non-initial according to whether that unit $\{\mathfrak{a}_i\}$ was earliest or not. Each empty set as unit we replace by itself (but write it $\overline{I}$). There results a regular expression. The state changes from $\overline{\mathfrak{a}}$ at time 1 to $\mathfrak{a}$ at time p, exactly if the event described by this regular expression occurs over the time $1, \ldots, p$. □

Corollary *The event represented by each of the following is likewise regular: an internal state at time $p + 1$, a property of the state at time p, a property of the internal state at time $p + 1$, a property of the internal state at time $p + s$ for an $s > 1$ which does not depend on the input over the time $p + 1, \ldots, p + s - 1$.*

Proof An event represented by a property of the state at time p is the disjunction of the events represented at time p by the states which have that property. The other modes of representation reduce to this via Lemma 1 in 5.3 (which applies here just as to McCulloch-Pitts nerve nets). □

Discussion The regular expressions obtained by the proof of Theorem 5 have only initial units or $\overline{I}$ as earliest units and are built of units of length 1 (and likewise after simplification by Lemma 2). It is clear in many examples that great simplifications can be obtained by use of equivalences (7.1); but we have made no study of the possibilities for proceeding systematically with such simplifications.

The study of the structure of a set of objects $\mathfrak{a}_1, \ldots, \mathfrak{a}_r$ under a binary relation R, which is at the heart of the above proof, might profitably draw on some algebraic theory.

It is of course essential to our arguments that the number of cells and the number of states for each be finite, so that the number of complete states is fixed in advance. A machine of Turing [TUR 36] is not a finite automaton, if the tape is considered as part of the machine, since, although only a finite number of squares of the tape are printed upon at any moment, there is no preassigned bound to this number. If the tape is considered as part of the environment, a Turing machine is a finite automaton which can in addition store information in the environment and reach for it later, so that the present input is not entirely independent of the past. Whether this comparison may lead to any useful insights into Turing machines or finite automata remains undetermined.

APPENDICES

10 Representability in a Finite Automaton with an Infinite Past

Theorem 6 *An event E is representable by a property of the state at time p of a finite automaton with an infinite past, only if E is definite.*

Proof With $k > 0$ input cells, a complete input is generated by choosing between the finite number 2^k of possible inputs at time p, then between the same number of possible inputs at time $p-1$, etc. ad infinitum.

By a theorem of Brouwer [BRO 24],[3] also given by König [KÖN 27], if

[3]Brouwer's treatment is intended for readers acquainted with intuitionistic set theory, and his main effort is to demonstrate the theorem intuitionistically.

for each input it is determined at some finite stage (i.e., from only the part of the input occupying the time $p, \ldots, p-u$ for some $u \geq 0$) whether an event occurs or not, then there is a number $n \geq 0$ such that for any input whether or not the event occurs is determined from only the part of it occupying the time $p, \ldots, p-n$. In this case the event would be definite of length $n+1$.

Now consider an indefinite event E. Contraposing Brouwer's theorem, there is an input $\mathtt{c}_0\mathtt{c}_1\mathtt{c}_2\ldots$ such that for every $u \geq 0$ it is not determined by the part of it $\mathtt{c}_0 \ldots \mathtt{c}_u$ for the time $p, \ldots, p-u$ whether E occurs or not.

Case 1: E does not occur for the input $\mathtt{c}_0\mathtt{c}_1\mathtt{c}_2\ldots$. Then for each u there is an input $\mathtt{c}_0^u\mathtt{c}_1^u\mathtt{c}_2^u\ldots$, coinciding with $\mathtt{c}_0\mathtt{c}_1\mathtt{c}_2\ldots$ over the time $p, \ldots, p-u$ and diverging from it at some earlier moment, for which E occurs.

Suppose E is represented by a property of the state a time p. Say the states which have the property are $\mathtt{a}_1, \ldots, \mathtt{a}_{r_1}$ and those which do not are $\mathtt{a}_{r_1+1}, \ldots, \mathtt{a}_r$.

Let S be the set of all the sequences of states $\mathtt{d}_0\mathtt{d}_1\mathtt{d}_2\ldots$ compatible with the present state being one of $\mathtt{a}_1, \ldots, \mathtt{a}_{r_1}$; i.e., $\mathtt{d}_0$ is one of $\mathtt{a}_1, \ldots, \mathtt{a}_{r_1}$, and each $\mathtt{d}_i$ has as its internal part that which is determined by $\mathtt{d}_{i+1}$ being the state at the immediately preceding moment. There are r_1 choices for $\mathtt{d}_0$, at most r for $\mathtt{d}_1$, at most r for $\mathtt{d}_2$, etc.

Any sequence of states $\mathtt{d}_0\mathtt{d}_1\mathtt{d}_2\ldots$ which can be assumed for the input $\mathtt{c}_0^u\mathtt{c}_1^u\mathtt{c}_2^u\ldots$ must belong to S, since E occurs for $\mathtt{c}_0^u\mathtt{c}_1^u\mathtt{c}_2^u\ldots$, and must in its first $u+1$ choices $\mathtt{d}_0 \ldots \mathtt{d}_u$ be compatible with $\mathtt{c}_0\mathtt{c}_1\mathtt{c}_2\ldots$, i.e., the external part of $\mathtt{d}_0 \ldots \mathtt{d}_u$ must be the input $\mathtt{c}_0 \ldots \mathtt{c}_u$ over the last $u+1$ moments $p, \ldots, p-u$ in $\mathtt{c}_0\mathtt{c}_1\mathtt{c}_2\ldots$.

By Brouwer's theorem, if for each sequence $\mathtt{d}_0\mathtt{d}_1\mathtt{d}_2\ldots$ belonging to S there were a u such that $\mathtt{d}_0 \ldots \mathtt{d}_u$ is incompatible with $\mathtt{c}_0\mathtt{c}_1\mathtt{c}_2\ldots$, there would be an n such that for each $\mathtt{d}_0\mathtt{d}_1\mathtt{d}_2\ldots$ belonging to S the part $\mathtt{d}_0 \ldots \mathtt{d}_n$ is incompatible with $\mathtt{c}_0\mathtt{c}_1\mathtt{c}_2\ldots$, contradicting the preceding remark for $u \geq n$.

So there is an infinite sequence $\mathtt{d}_0\mathtt{d}_1\mathtt{d}_2\ldots$ in S which is compatible with $\mathtt{c}_0\mathtt{c}_1\mathtt{c}_2\ldots$. But $\mathtt{d}_0$ is one of the states $\mathtt{a}_1, \ldots, \mathtt{a}_{r_1}$, although E does not occur for $\mathtt{c}_0\mathtt{c}_1\mathtt{c}_2\ldots$, contrary to our supposition that E is represented by the state at time p being one of $\mathtt{a}_1, \ldots, \mathtt{a}_{r_1}$.

Case 2: E occurs for the input $\mathtt{c}_0\mathtt{c}_1\mathtt{c}_2\ldots$. Applying to $\overline{E}$ the reasoning applied in Case 2 to E, it is absurd that $\overline{E}$, and hence that E, be represented by a property of the state at time p.

□

11 Representability with a Finite Past but an Arbitrary Initial Internal State

Theorem 7 *An event E is representable by a property of the state at time p of a finite automaton started with an arbitrary internal state at time* 1, *only if E is non-initial definite of length* 1.

Proof Let E be an event not non-initial definite of length 1. Then there is some input $\mathfrak{c}$ for the moment p such that whether or not E occurs is not determined by $\mathfrak{c}$ alone; i.e., different choices $\mathfrak{c}'_1 \dots \mathfrak{c}'_{p'-1}$ and $\mathfrak{c}''_1 \dots \mathfrak{c}''_{p''-1}$ of the input over $1, \dots, p-1$ for $p = p'$ and $p = p''$ together with $\mathfrak{c}$ at p make E occur or not occur, respectively. Suppose E is represented by a certain property of the state at time p for a given initial internal state $\mathfrak{b}_1$. Consider the inner states $\mathfrak{b}'$ and $\mathfrak{b}''$ produced at times p' and p'' from the initial internal state $\mathfrak{b}_1$ by the inputs $\mathfrak{c}'_1 \dots \mathfrak{c}'_{p'-1}$ and $\mathfrak{c}''_1 \dots \mathfrak{c}''_{p''-1}$, respectively. Now at time 1 let the input be $\mathfrak{c}$ and the internal state be $\mathfrak{b}'$ or $\mathfrak{b}''$, respectively. Then the property of the state is possessed or not possessed, respectively. Thus the property cannot represent E for both $\mathfrak{b}'$ and $\mathfrak{b}''$ as initial internal state; for one of them it gives a false result for $p = 1$ and $\mathfrak{c}$ as input. □

12 Primitive Recursiveness of Regular Events

To illustrate that only logical and mathematical symbolism on the level of number theory is necessary to express regular events, we state the following theorem. The notion of relative primitive recursiveness is defined in [KLE 52]. For conformity with the notation there, the time variables for this theorem shall range over $0, 1, 2, \dots$ instead of $1, 2, 3, \dots$.

Theorem 8 *For any regular event E referring to input neurons $\mathcal{N}_1, \dots, \mathcal{N}_k$, the predicate $E(p)$ ($\equiv$ E occurs ending with time p) is primitive recursive in the predicates $N_1(t), \dots, N_k(t)$.*

Method of proof Using Theorem 3, $E(p)$ is equivalent to the existence of a certain kind of a string of states $\mathfrak{d}_p \dots \mathfrak{d}_0$ (cf. Section 9). □

13 A Simple Example of an Irregular Event

Consider the event E described as follows: $\mathcal{N}$ fired at time u^2 for every u such that $u^2 \leq p$ and only at those times. In symbols, $E(p) \equiv (t)_{t \leq p}[N(t) \equiv (Eu)_{u \leq p} t = u^2]$.

No finite automaton can represent E, and hence by Theorem 3 E is not regular. For suppose E is represented by a property of the state at time p of

a finite automaton (admitting states $\mathfrak{a}_1, \ldots, \mathfrak{a}_r$); say the states which have this property are $\mathfrak{a}_1, \ldots, \mathfrak{a}_r$.

Consider any number s such that $2s > r_1$. Suppose $\mathcal{N}$ fires at times $1, 4, 9, \ldots, s^2$ and never thereafter. Then E occurs for $p = 1, 2, \ldots, s^2 + 2s$ $(= (s+1)^2 - 1)$ and for no greater p.

Consider the states $\mathfrak{d}_1, \mathfrak{d}_2, \mathfrak{d}_3, \ldots$ of the automaton at the times $s^2 + 1, s^2 + 2, s^2 + 3, \ldots$. Beginning with time $s^2 + 1$, $\mathcal{N}$ never fires, so the external state is constant. Thus each of the states $\mathfrak{d}_1, \mathfrak{d}_2, \mathfrak{d}_3, \ldots$ after the first is determined by the immediately preceding one. So, since there are only r states altogether, the sequence $\mathfrak{d}_1, \mathfrak{d}_2, \mathfrak{d}_3, \ldots$ is ultimately periodic.

However, during the time $s^2 + 1, \ldots, s^2 + 2s$ the state must be one of $\mathfrak{a}_1, \ldots, \mathfrak{a}_{r_1}$, since E occurs for these values of p. Hence, since $2s > r_1$, the period must already have become established (i.e., the first repetition in $\mathfrak{d}_1, \mathfrak{d}_2, \mathfrak{d}_3, \ldots$ must already have occurred) by the time $s^2 + 2s$. Hence the state at time $(s+1)^2$ is one of $\mathfrak{a}_1, \ldots, \mathfrak{a}_{r_1}$, although E does not occur for $p = (s+1)^2$.

It is not suggested that the event would be of any biological significance. The example is given to show the mathematical limitations to what events can be represented.

References

[BRO 24] BROUWER, L. E. J., *Beweis, dass jede volle Funktion gleichmässig stetig ist.* Verhandelingen Koninklijke Nederlandsche Akademie van Wetenschappen, Amsterdam, vol. 27 (1924), pp. 189–193. Another version: *Uber Definitionsbereiche von Funktionen*, Mathematische Annalen, vol. 97 (1927), pp. 60–75.

[GER 53] GERARD, Ralph W., *What is memory?* Scientific American, vol. 189, no. 3, September 1953, pp. 118–126.

[HIL 28] HILBERT, David and ACKERMANN, Wilhelm, *Grundzüge der theoretischen Logik.* First ed., Berlin (Springer) 1928, viii + 120 pp. Third ed., Berlin, Göttingen, Heidelberg (Springer) 1949, viii + 155 pp. Eng. tr. of the second ed., *Principles of mathematical logic*, New York (Chelsea) 1950, xii + 172 pp.

[HOU 45] HOUSEHOLDER, A. S. and LANDAHL, H. D., *Mathematical biophysics of the central nervous system.* Mathematical biophysics monograph series, no. 1, Bloomington, Indiana (Principia press) 1945, ix + 124 pp.

[KLE 52] KLEENE, S. C., *Introduction to metamathematics.* Amsterdam (North Holland Pub. Co.), Groningen (Noordhoff) and New York (Van Nostrand) 1952, x + 550 pp.

[KÖN 27] KÖNIG, D., *Über eine Schlussweise aus dem Endlichen ins Unendliche.* Acta litterarum ac scientiarum (Szeged), Sect. math., vol. III/II (1927), pp. 121–130 (particularly the appendix pp. 129–130).

[McC 49] McCULLOCH, Warren S., *The brain as a computing machine.* Electrical engineering, vol. 68 (1949), pp. 492–497.

[McC 43] McCULLOCH, Warren S. and PITTS, Walter, *A logical calculus of the ideas immanent in nervous activity.* Bulletin of mathematical biophysics, vol. 5 (1943), pp. 115–133.

[QUI 52] QUINE, W. V., *The problem of simplifying truth functions.* American mathematical monthly, vol. 59 (1952), pp. 521–531.

[TUR 36] TURING, A. M., *On computable numbers, with an application to the Entscheidungsproblem.* Proceedings of the London Mathematical Society, ser. 2, vol. 42 (1936-7), pp. 230–265. *A correction*, ibid., vol. 43 (1937), pp. 544–546.