GEDANKEN-EXPERIMENTS ON SEQUENTIAL MACHINES

Edward F. Moore

INTRODUCTION

This paper is concerned with finite automata[1] from the experimental point of view. This does not mean that it reports the results of any experimentation on actual physical models, but rather it is concerned with what kinds of conclusions about the internal conditions of a finite machine it is possible to draw from external experiments. To emphasize the conceptual nature of these experiments, the word "gedanken-experiments" has been borrowed from the physicists for the title.

The sequential machines considered have a finite number of states, a finite number of possible input symbols, and a finite number of possible output symbols. The behavior of these machines is strictly deterministic (i.e., no random elements are permitted in the machines) in that the present state of a machine depends only on its previous input and previous state, and the present output depends only on the present state.

The point of view of this paper might also be extended to probabilistic machines (such as the noisy discrete channel of communication theory[2]), but this will not be attempted here.

EXPERIMENTS

There will be two kinds of experiments considered in this paper. The first of these, called a simple experiment, is depicted in Figure 1.

[1] The term "finite" is used to distinguish these automata from Turing machines [considered in Turing's "On Computable Numbers, with an Application to the Entscheidungsproblem", <u>Proc</u>. <u>Lond</u>. <u>Math</u>. <u>Soc</u>., (1936) Vol. 24, pp. 230-265] which have an infinite tape, permitting them to have more complicated behavior than these automata.

[2] Defined in Shannon's "A Mathematical Theory of Communication", B.S.T.J. Vol. 27, p. 406.

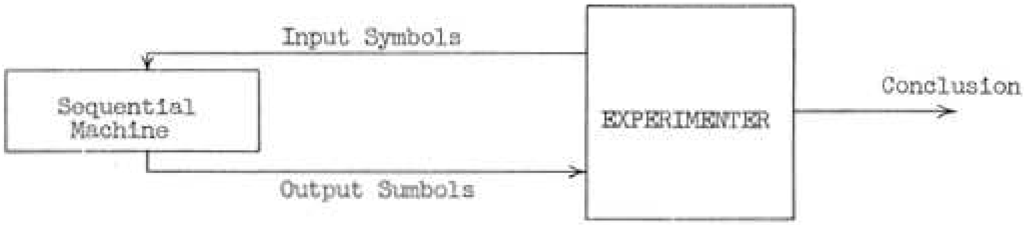

FIGURE 1. Schematic Diagram of a Simple Experiment

A copy of the sequential machine being observed experimentally will receive successively certain input symbols from the experimenter. The sequence of output symbols will depend on the sequence of input symbols (the fact that the correspondence is between sequences rather than individual symbols is responsible for the terminology "sequential machine") in a way that depends on which particular sequential machine is present and its initial state.

The experimenter will choose which finite sequence of input symbols to put into the machine, either a fixed sequence, or one in which each symbol depends on the previous output symbols. This sequence of input symbols, together with the sequence of output symbols, will be called the outcome of the experiment. In addition there can be a conclusion which the experimenter emits, the exact nature of which need not be specified. The conclusion might be thought of as a message typed out on a typewriter, such as "The machine being experimented on was in state q_1 at the beginning of the experiment". It is required that the conclusion depend only on which experiment is being performed and what the sequence of output symbols was.

The second kind of experiment considered in this paper is the multiple experiment, shown in Figure 2.

In this case the experimenter has access to several copies of the same machine, each of which is initially in the same state. The experimenter can send different sequences of inputs to each of these K copies, and receive from each the corresponding output sequence.

In each of these two kinds of experiments the experimenter may be thought of as a human being who is trying to learn the answer to some question about the nature of the machine or its initial state. This is not the only kind of experimenter we might imagine in application of this theory; in particular the experimenter might be another machine. One of the problems we consider is that of giving explicit instructions for performing the experiments, and in any case for which this problem is completely solved it is possible to build a machine which could perform the experiment.

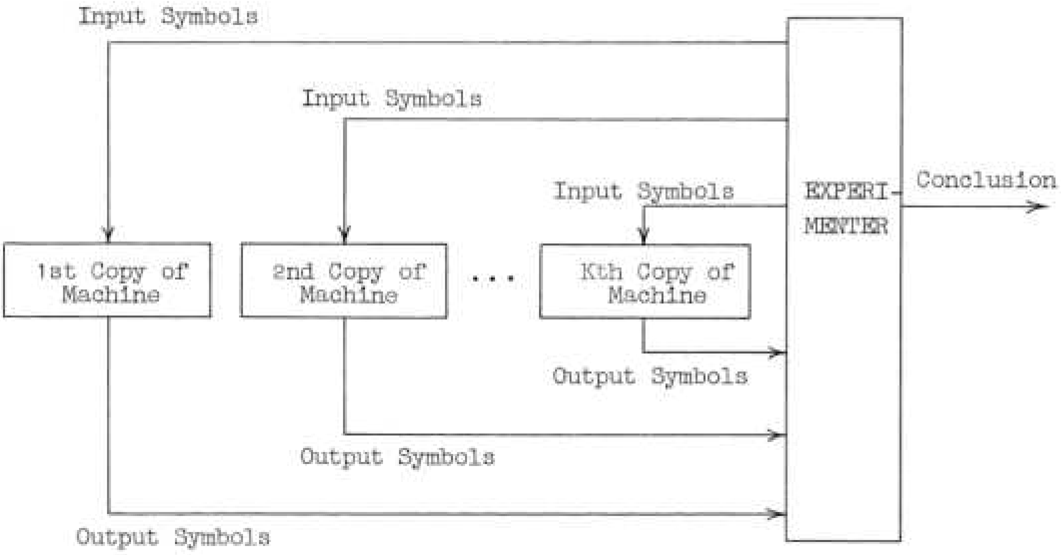

FIGURE 2. Schematic Diagram of a Multiple Experiment

EXAMPLES

It may be instructive to consider several situations for which this sort of theory might serve as a mathematical model.

The first example is one in which one or more copies of some secret device are captured or stolen from an enemy in wartime. The experimenter's job is to determine in detail what the device does and how it works. He may have partial information, e.g., that it is a bomb fuze or a cryptographic device, but its exact nature is to be determined. There is one special situation that can occur in such an experiment that is worthy of note. The device being experimented on may explode, particularly if it is a bomb, a mine, or some other infernal machine. Since the experimenter is presumably intelligent enough to have anticipated this possibility, he may be assumed to have conducted his experimentation by remote control from a safe distance. However, the bomb or mine is then destroyed, and nothing further can be learned from it by experimentation. It is interesting to note that this situation can be represented exactly by the theory. The machine will have some special state q_n, the exploded state. The transitions defining the machine will be such that there exists a sequence of inputs that can cause the machine to go into state q_n, but no input which will cause it to leave the state. Hence, if the experimenter happens to give the wrong sequence to the machine, he will be unable to learn anything further from this copy of the machine.

There is a somewhat artificial restriction that will be imposed on the action of the experimenter. He is not allowed to open up the machine and look at the parts to see what they are and how they are interconnected. In this military situation, such a restriction might correspond to the machine being booby trapped so as to destroy itself if tampered with. It might also correspond to an instance where the components are so unfamiliar that nothing can be gained by looking at them. At any rate, we will always impose this somewhat artificial restriction that the machines under consideration are always just what are sometimes called "black boxes", described in terms of their inputs and outputs, but no internal construction information can be gained.

Another application might occur during the course of the design of actual automata. Suppose an engineer has gone far enough in the design of some machine intended as a part of a digital computer, telephone central office, automatic elevator control, etc., to have described his machine in terms of the list of states and transitions between them, as used in this paper. He may then wish to perform some gedanken-experiments on his intended machine. If he can find, for instance, that there is no experimental way of distinguishing his design from some machine with fewer states, he might as well build the simpler machine.

It should be remarked that from this engineering point of view certain results closely paralleling parts of this paper (notably the reduction described in Theorem 4) have recently been independently found by D. A. Huffman in his Ph.D. thesis in Electrical Engineering (M.I.T.). His results are to appear in the Journal of the Franklin Institute.

Still another situation of which this theory is a mathematical model occurs in the case of the psychiatrist, who experiments on a patient. He gives the patient inputs (mainly verbal), and notes the outputs (again mainly verbal), using them to learn what is wrong with the patient. The black box restriction corresponds approximately to the distinction between the psychiatrist and the brain surgeon.

Finally, another situation of which this might conceivably be a mathematical model occurs when a scientist of any sort performs an experiment. In physics, chemistry, or almost any other science the inputs which an experimenter puts into his experiment and the outputs he gets from it do not correspond exactly to the things the experimenter wishes to learn by performing the experiment. The experimenter is frequently forced to ask his questions in indirect form, because of restrictions imposed by intractable laws of nature. These restrictions are somewhat similar in their effect on the organization of the experiment to the black box restriction.

The analogy between this theory and such scientific experimentation is not as good as in the previous situations, because actual experiments may be continuous and probabilistic (rather than finite and deterministic), and also because the experiment may not be completely isolated from the experimenter, i.e., the experimenter may be experimenting on a system of which he himself is a part. However, certain qualitative results of the theory may be of interest to those who like to speculate about the basic problems of experimental science.

CONVENTIONS

Each machine will have a finite number n of states, which will be called $q_1, q_2, \ldots, q_n$ a finite number m of possible input symbols which will be called $S_1, S_2, \ldots, S_m$, and a finite number p of possible output symbols, which will be called $S_{m+1}, S_{m+2}, \ldots, S_{m+p}$. In several examples used in this paper we will have $m = 2$, $p = 2$, $S_1 = S_3 = 0$, and $S_2 = S_4 = 1$.

Time is assumed to come in discrete steps, so the machine can be thought of as a synchronous device. Since many of the component parts of actual automata are variable in their speed, this assumption means the theory has not been stated in the most general terms. In practice, some digital computers and most telephone central offices have been designed asynchronously. However, by providing a central "clock" source of uniform time intervals it is possible to organize even asynchronous components so that they act in the discrete time steps of a synchronous machine. Digital computers and other electronic automata are usually built in this synchronous fashion. The synchronous convention is used in this paper since it permits simpler exposition, but the fact that these results can be translated with very little change into asynchronous terms should be obvious from the fact that Huffman wrote his paper in terms of the asychronous case.

The state that the machine will be in at a given time depends only on its state at the previous time and the previous input symbol. The output symbol at a given time depends only on the current state of the machine. A table used to give these transitions and outputs will be used as the definition of a machine. To illustrate these conventions, let us consider the following example of a machine:

Machine A

Present State		
Previous State	Previous Input	
	0	1
q_1	q_4	q_3
q_2	q_1	q_3
q_3	q_4	q_4
q_4	q_2	q_2

Present State	Present Output
q_1	0
q_2	0
q_3	0
q_4	1

These two tables give the complete definition of a machine (labelled machine A, for future reference). In the left table, the present state of the machine is given as a function of the previous state and the previous input. In the right table, the present output of the machine is given as a function of the present state.

An alternate way of representing the description of a machine can also be used, which may be somewhat more convenient to follow. This other representation, called a transition diagram, consists of a graph whose vertices correspond to the states of the machine represented, and whose edges correspond to the possible transitions between those states. Each vertex of this transition diagram will be drawn as a small circle, in which is written the symbol for the corresponding state, a semicolon, and the output which the machine gives in that state.

Each pair of these circles will be joined by a line if there is a direct transition possible between the corresponding pair of states. An arrowhead will point in the direction of the transition. Beside each such line there will be written a list of the possible input symbols which can cause the transition. Below is given a transition diagram for machine A:

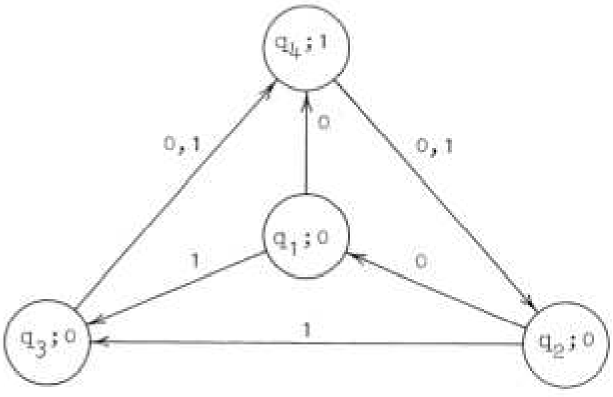

FIGURE 3. Transition Diagram of Machine A

An experiment can be performed on this machine by giving it some particular sequence of inputs. As an example, the sequence 000100010 might be used. If the machine is initially in the state q_1, the outcome of this experiment would be:

$$\begin{array}{ccccccccc}
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
q_1 & q_4 & q_2 & q_1 & q_3 & q_4 & q_2 & q_1 & q_3 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0
\end{array}$$

where the first line of the above is the sequence of inputs, the second line is the sequence of states, and the third line is the sequence of outputs. The last two lines can be obtained from the first by use of the tabular definition of machine A or its transition diagram. It should be emphasized that only the bottom line of the above is observable by the experimenter, and the sequence of states is hidden away, usable only in arriving at or explaining the observable results of the experiment.

Suppose that the same sequence of inputs mentioned above is presented to machine A, initially in some other state. The outcome of the experiment would be one of the following, according as the initial state is q_2, q_3, or q_4:

$$\begin{array}{ccccccccc}
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
q_2 & q_1 & q_4 & q_2 & q_3 & q_4 & q_2 & q_1 & q_3 \\
0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0
\end{array}$$

$$\begin{array}{ccccccccc}
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
q_3 & q_4 & q_2 & q_1 & q_3 & q_4 & q_2 & q_1 & q_3 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0
\end{array}$$

$$\begin{array}{ccccccccc}
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
q_4 & q_2 & q_1 & q_4 & q_2 & q_1 & q_4 & q_2 & q_3 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0
\end{array}$$

Even though this example of an experiment involved putting a predetermined sequence of input symbols into the machine, it should not be assumed that this is the only kind of experiment permitted. In general, the inputs to the machine can depend on its previous outputs, permitting the course of the experiment to branch.

There would be several ways of specifying such a branching experiment, but for the purposes of this paper, a loose verbal description of such experiments will be used. If it were desired to make these descriptions more formal, the experimenter could be described as another sequential machine, also specified in terms of its internal states, inputs, and outputs. The output of the machine being experimented on would serve as input to the experimenter and vice versa. The experimenter would also have another output in which it would summarize the results of the experiment, indicating what has been learned about the machine by the experiment.

In the simple sequence given above as an example of an experiment, it is natural to define the length of the experiment as 9, since this is the number of terms in the input sequence, and the number of discrete steps of time required to perform this experiment. But in the case of an experiment with possible branches during its performance, some of these branches may lead to a conclusion more quickly than others. In this case the length required for the longest possible alternative would be taken as the length of the experiment.

Although a branching experiment is the most general type of deterministic experiment, most of the experiments which will be required in the proofs of this paper can simply be sequences. For example, the shortest simple experiment which can be used to distinguish between two states (of the same or different machines) is merely a sequence. For if this is the shortest experiment, the result is not known until the last step, i.e., the output sequences coming to the experimenter are the same except for the last term. This term comes too late to affect any part of the experiment.

Two machines, S and T, will be said to be isomorphic if the table describing S can be obtained from the table describing T by substituting new names for the states wherever they occur as either the arguments or the entries of the table. Clearly, isomorphic machines will always have the same behavior, and will be indistinguishable from one another by any experiment.

Since distinguishability has already been referred to several times, and is vital to every proof in this paper, it will be explained in some detail.

A state q_i of a machine S will be said to be indistinguishable from a state q_j of S if and only if every experiment performed on S starting in state q_i produces the same outcome as it would starting in state q_j.

A pair of states will be said to be distinguishable if they are not indistinguishable. Hence, q_i is indistinguishable from q_j if and only if there exists some experiment of which the outcome depends on which of these two states S was in at the beginning of the experiment.

Similarly, we can say that a state q_i of a machine S is distinguishable (or indistinguishable) from a state q_j of a machine T if there exists an experiment (or there does not exist an experiment) of which the outcome starting with machine S in state q_i differs from the outcome starting with machine T in state q_j.

Finally, distinguishability and indistinguishability can be defined for pairs of machines. A machine S will be said to be distinguishable from a machine T if and only if at least one of the following two conditions hold: either

(1) there exists some state q_i of S, and some experiment the outcome of which beginning with S in state q_i differs from its outcome beginning with T in each of its states, or

(2) there exists some state q_j of T, and some experiment the outcome of which beginning with T in state q_j differs from its outcome beginning with S in each of its states.

S and T will be said to be indistinguishable if and only if they are not distinguishable, or, in other words, if both of the following two conditions hold:

(1) for every state q_i of S, and every experiment, there exists a state q_j of T such that the experiment beginning with machine S in state q_i produces the same outcome as the experiment beginning with machine T in state q_j, and

(2) for every state q_j of T, and every experiment, there exists a state q_i of S such that the experiment beginning with machine T in state q_j produces the same outcome as the experiment beginning with machine S in state q_i.

If S is indistinguishable from T, then the two machines are alike in their behavior (although they may differ in their structure), and may be thought of as being interchangeable. In any practical application of real machines, the manufacturer can take advantage of this equivalence, and produce whichever of the two machines is cheaper to build, easier to repair, or has some other desirable internal property.

Distinguishability and indistinguishability are defined here as binary relations. That is, they hold between a pair of machines or a pair of states. This does not mean that an experiment which distinguishes between them must be a multiple experiment. In many cases a simple experiment suffices. In any event, we perform the experiment on just one of the two machines or states we wish to distinguish, and its outcome depends on which of the two was present. In these cases we may think of the conclusion which the experimenter reaches as being of the form: "If the machine being examined was either S or T, then it is now known to be T."

This is certainly an extremely elementary kind of a conclusion, which makes a binary choice between two alternatives. Part of this paper will deal with methods of building up more complicated conclusions from such elementary ones.

An obvious modification of distinguishability is to state whether the machines which can be distinguished require multiple experiments to tell them apart or not. In the case of pairs of states, the two kinds of distinguishability can easily be seen to coincide.

In the course of the proofs given below, it will frequently be convenient to look at experiments in terms of what is actually happening inside the machines. Although the experimenters are not permitted to look inside the black boxes, we are under no such restriction. In fact, we will be able to learn more about the limitations imposed by the black box restriction if we have no such restriction on our observations, constructions, or proofs.

AN ANALOGUE OF THE UNCERTAINTY PRINCIPLE

The first theorem to be proved will be concerned with an interesting qualitative property of machines.

Theorem 1: There exists a machine such that any pair of its states are distinguishable, but there is no simple experiment which can determine what state the machine was in at the beginning of the experiment.

The machine A, already described on the previous pages, satisfies the conditions of the theorem. The previously described experiment will distinguish between any pair of states, except the pair (q_1, q_3). That is, given any other pair of states, if it is known that the machine is in one state of this pair at the beginning of the experiment, applying this experiment will give an output that depends on which state the machine was in. In order to distinguish between q_1 and q_3, the experiment should consist of applying the sequence 11. The outcome of this will be:

$$\begin{array}{cc} 1 & 1 \\ q_1 & q_3 \\ 0 & 0 \end{array} \qquad\qquad \begin{array}{cc} 1 & 1 \\ q_3 & q_4 \\ 0 & 1 \end{array}$$

Thus there exists a simple experiment which can distinguish between any pair of states. Furthermore, the multiple experiment which uses two copies of the machine, sending one of the two previously mentioned sequences to each, can obtain enough information to completely specify what state the machine was in at the beginning of the experiment.

To complete the proof, it need only be shown that given only one copy of machine A, there is no experiment which can determine whether it was in state q_1 at the beginning of the experiment.

It is clear that any experiment will distinguish between q_1 and q_4, since the first output symbol will be different. But any simple experiment that distinguishes q_1 from q_2, cannot distinguish q_1 from q_3. To see this, note that any experiment which begins with the input 1 does not permit q_1 to be distinguished from q_2 (since in either case the first output is 0 and the second state is q_3, so that no future inputs can produce different outputs). Similarly any experiment which begins with the input 0 does not permit q_1 to be distinguished from q_3.

This result can be thought of as being a discrete-valued analogue of the Heisenberg uncertainty principle. To point out the parallel, both the uncertainty principle and this theorem will be restated in similar language.

The state of an electron E will be considered specified if both its velocity and its position are known. Experiments can be performed which will answer either of the following:

(1) What was the position of E at the beginning of the experiment?

(2) What was the velocity of E at the beginning of the experiment?

In the case of machine A, experiments can be performed which will answer either of the following:

(1) Was A in state q_2 at the beginning of the experiment?

(2) Was A in state q_3 at the beginning of the experiment?

In either case, performing the experiment to answer question 1 changes the state of the system, so that the answer to question 2 cannot be obtained. In other words, it is only possible to gain partial information about the previous history of the system, since performing experiments causes the system to "forget" about its past.

By analogy with the uncertainty principle, could we also state that the future state of machine A cannot be predicted from past experimental results? Here the analogy ends. Even though we cannot learn by experiment what state machine A was in at the beginning of the experiment, we can learn what state it is in at the end of the experiment. In fact, at the end of the first experiment described, machine A will be in one particular predetermined state (independent of its initial state), namely the state q_3.

Despite the incompleteness of the analogy, it does seem interesting that there is an analogue of the uncertainty principle in this discrete, deterministic system. Any applications of this example to causality, free will, or other metaphysical problems will be left to the reader.

FURTHER THEOREMS ON DISTINGUISHABILITY

Theorem 2: Given any machine S and any multiple experiment performed on S, there exist other machines experimentally distinguishable from S for which the original experiment would have had the same outcome.

Let S have n states $q_1, \ldots q_n$, and let the experiment have length k. Then define a machine T having $n(k+1)$ states $q_1, q_2, \ldots, q_{n(k+1)}$ as follows:

If the machine S goes from state q_i to q_j when it receives the input symbol a, then let T go from q_{i+tn} to $q_{j+(t+1)n}$ under the same input, for all t such that $0 \leq t < k$, but let T go from q_{i+kn} to q_{j+kn}.

If the machine S has output symbol b in state q_i, let T have output symbol b in state q_{i+tn}, for $0 \leq t < k$, but let T have some output symbol different from b in state q_{i+kn}.

Then at the step $t+1$ of any simple experiment, the machine T will be in state q_{i+tn} whenever machine S is in state q_i and $0 \leq t < k$. But at any step later than the kth, machine T will be in state q_{i+kn}. Thus it can be seen that for the first k steps of any simple experiment, the outputs of S and T will be alike. But after the kth step, the outputs of S and T will always be different. The extension to multiple experiments is immediate.

This result means that it will never be possible to perform experiments on a completely unknown machine which will suffice to identify it from among the class of all sequential machines. If, however, we restrict the class to be a smaller one, it may be possible. In particular, much of the rest of this paper will be concerned with the case where the class consists of all machines with n states or fewer, m input symbols or fewer, and p output symbols or fewer. Such a machine will be called an (n, m, p) machine.

Definition: A machine S will be said to be strongly connected if for any ordered pair (q_i, q_j) of states of S, there exists a sequence of inputs which will take the machine from state q_i to state q_j.

The term "strongly connected" is used since any such machine will have a transition diagram which is a connected graph, but the converse is not true. A counter-example to the converse is given by the following machine:

Machine B

Present State		
Previous State	Previous Input 0	1
q_1	q_2	q_3
q_2	q_1	q_3
q_3	q_1	q_2
q_4	q_2	q_2

Present State	Present Output
q_1	0
q_2	1
q_3	0
q_4	0

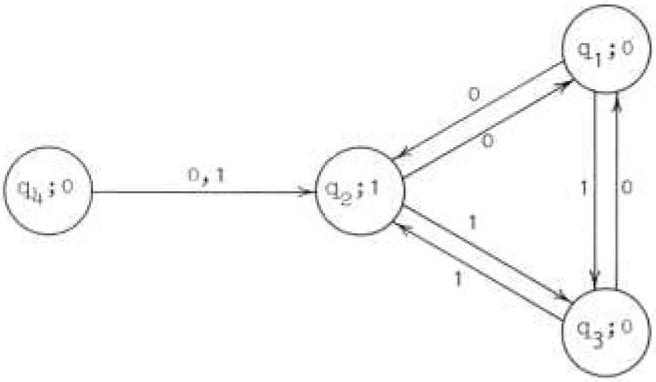

FIGURE 4. Transition Diagram of Machine B

<u>Theorem 3</u>: If S is a strongly connected machine, and T is indistinguishable from S by any simple experiment, then for every state q_i of S there exists a state q_j of T which is indistinguishable from q_i by any simple experiment.

Since T is indistinguishable from S by any simple experiment, we have, as one of the two conditions implied by indistinguishability, that given any state q_i of S, and any simple experiment on S beginning in state q_i, there exists a corresponding state q_j of T such that the same experiment, starting with a copy of T in state q_j, will produce the same sequence of output symbols. This theorem states that if S is strongly connected, q_j can be chosen independently of the experiment. That is, q_j corresponds to q_i for all experiments, rather than just this particular experiment.

To prove the theorem first note that if we consider an experiment consisting of any sequence of input symbols applied to machine S in state q_i, there must have been states of T which would have given the same sequence of outputs. With each such sequence of input symbols, we associate the set of states that machine T could be in at the end of this sequence after having produced the same sequence of outputs that S would produce starting in state q_1. Then if we lengthen the sequence of input symbols, the number of elements in the associated set can only decrease, but it can never become zero (or else this would give an experiment which distinguishes S from T).

Hence, we can choose a particular sequence and extend it until the number of elements in the associated set of states of T can no longer be decreased by any further extension. Then we add to this sequence a further sequence which will cause machine S to go successively into every one of its states at least once, if the entire sequence is applied starting in state q_1.

Then for each state q_i of S, consider the set Y of states which are associated with the subsequence obtained by truncating the original sequence at the last time it causes S to go into state q_i. Then Y is non-empty, and every member is indistinguishable from q_i. This follows from the fact that if q_j is a member of Y and is distinguishable from q_i, the experiment that distinguishes them defines a sequence, which when added to the truncated sequence above, would give a further reduction of the number of elements in its associated set. But this contradicts the definition of the original sequence.

Note the words "strongly connected" cannot be removed from the statement of Theorem 3. A counter-example is given by machine B, defined just before Theorem 3, which is indistinguishable by any simple experiment from the machine B', defined by removing the bottom row from each of the two tables that define machine B. However, the state q_4 of machine B is distinguishable from every state of B'.

Theorem 4: The class of all machines which are indistinguishable from a given strongly connected machine S by any simple experiment has a unique (up to an isomorphism) member with a minimal number of states. This unique machine, called the reduced form of S, is strongly connected, and also has the property that any two of its states are distinguishable.

Given any machine T, indistinguishable from S, define the relation R to hold between states of S and states of T if they are indistinguishable by a simple experiment. That is, the state q_1 of S will have the relation R to the state q_j of T if and only if there is no simple experiment which can distinguish them.

Then by Theorem 3 the domain of the relation R is the set of all states of S. And, after verifying the transitivity of indistinguishability it can be seen that any two states of S are indistinguishable from each other if and only if they are indistinguishable from the same state of T. Hence, the number of equivalence classes into which the states of S are partitioned by the equivalence relation of indistinguishability is the smallest number of states which T can have.

Let us define a machine T* with exactly this many states, associating each state with one such equivalence class. We can define the output symbol for each state of T* to be the output symbol for any state in its equivalence class, since if the states are indistinguishable, they must give the same output symbols. We define the transitions by letting state q_i of T* go into state q_j of T* upon receiving the input symbol a, if and only if some member of the equivalence class associated with q_i goes into some member of the equivalence class associated with q_j upon receiving the input symbol a. There is never any ambiguity in this definition, since indistinguishable states cannot have transitions which take them into distinguishable ones (or else this would give a way of distinguishing the original indistinguishable states).

Next, T* can be seen to be indistinguishable from S as an immediate consequence of its definition. Also T* is strongly connected, since to go between states q_i and q_j of T*, use the sequence which goes from any state in the equivalence class associated with q_i to any state in the equivalence class associated with q_j.

Then to show that T* is unique up to an isomorphism, consider any other machine T, having the same number of states, and also indistinguishable from S. Then since T will also be indistinguishable from T*, and T* is strongly connected we can apply Theorem 3. Then defining another relation R as done earlier in the proof, note that it can be seen to be a 1:1 correspondence between the states of the two machines, and in fact, it is the desired isomorphism.

Definition: A machine S will be said to be in reduced form, if and only if S is the reduced form of S.

Theorem 5: If S is a strongly connected machine, then S is in reduced form, if and only if any pair of its states are distinguishable. To prove the converse, consider the relation of indistinguishability as in the proof of Theorem 4: it partitions the states of S into equivalence classes, each having just one member. Hence, the reduced form of S as constructed above has exactly as many states as S, and the uniqueness of the reduced form of S completes the proof.

The following is an example of a machine which this theorem shows to be not in reduced form. This particular example has just one pair of states which are indistinguishable:

Machine C

Present State		
Previous State	Previous Input	
	0	1
q_1	q_2	q_1
q_2	q_3	q_1
q_3	q_2	q_1

Present State	Present Output
q_1	0
q_2	1
q_3	0

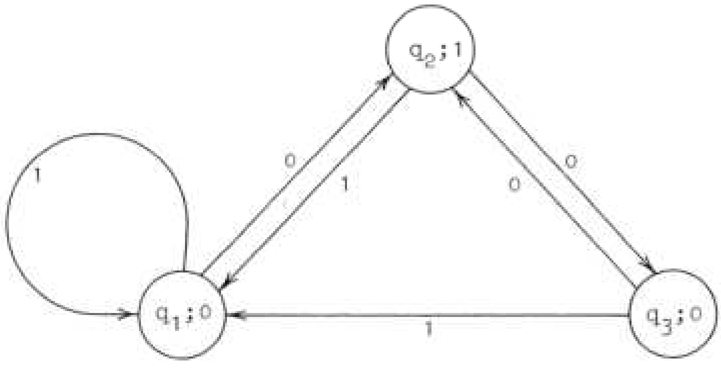

FIGURE 5. Transition Diagram of Machine C

In connection with these theorems, it might be mentioned that not every machine indistinguishable from a strongly connected machine is strongly connected. The machines B and B', previously described, also serve as an example of this.

However, since the reduced form of a machine is unique and has no indistinguishable states, it may be thought of as a simplified version of the machine, with all unessential parts of its description removed. The reduction of a machine to its reduced form is closely related to one of the steps proposed by D. A. Huffman as a step in the design of sequential machines.

The reduced form will be considered the natural form in which to describe a strongly connected machine, and the remaining theorems of this paper will be written in a form so as to apply directly to machines in reduced form. The indirect application of these results to other strongly connected machines is also sometimes possible.

THEOREMS CONCERNING LENGTHS OF EXPERIMENTS

The theorems proved heretofore have mainly been concerned with qualitative questions, i.e., whether or not it is possible to perform experiments which answer questions about the current state of a machine or its internal structure. The remaining theorems will be concerned with how many steps these experiments require, and their proofs will include methods for designing the experiments.

Theorem 6: If S is an (n, m, p) machine such that any two of its states are distinguishable, then they are distinguishable by a simple experiment of length $n-1$.

For each positive integer k, we define the relation R_k to hold between any two states q_i and q_j of S if and only if q_i is indistinguishable from q_j by any experiment of length k. Since each R_k can be seen to be an equivalence relation, it defines a partition P_k of the set Z of states of S into equivalence classes.

Then P_{k+1} is a refinement of P_k; that is, if two states are indistinguishable by any experiment of length $k+1$, they are indistinguishable by an experiment of length k. Further, if P_k does not subdivide Z into subsets having just one member, then P_{k+1} is a proper refinement of P_k. To show this, choose any two states q_i and q_j which are indistinguishable by an experiment of length k. Since by hypothesis they are distinguishable, consider the shortest sequence of inputs which will serve as an experiment to distinguish them. If this sequence of inputs is of length r, consider the pair of states which q_i and q_j are transformed into by the first $r-k-1$ inputs of this sequence. This pair of states is distinguishable by an experiment of length $k+1$ (namely, the rest of the above sequence) but not by any experiment of length k (for such an experiment would contradict the minimal length of the above sequence).

Since P_1 partitions Z into at least two subsets (for otherwise every state would have the same output associated, and hence no pairs of states are distinguishable) we can prove by induction from above that if $k \leq n - 1$, P_k partitions Z into at least $k+1$ subsets, which for the case $k = n - 1$ completes the proof of the theorem.

The above proof suggests a method for finding the shortest experiments for distinguishing between any two states. First construct P_1, by subdividing Z into sets of states giving the same output symbol. Then, proceeding by recursion, P_{k+1} can be constructed from P_k. If any two states q_i and q_j undergo transitions into states which belong to different classes of P_k upon receiving the same input symbol a, then q_i and q_j should be put into different classes of P_{k+1}, and a is the first symbol of an experiment for distinguishing between q_i and q_j in $k+1$ steps. If, however, under all input symbols q_i and q_j remain together in the same classes of P_k, they are indistinguishable by any experiment of length $k+1$, and hence belong in the same class of P_{k+1}. By continuing the recursion until any desired pair of states can be distinguished, this method constructs an experiment. It proceeds backwards; that is, the last step of the experiment is found first, and at the end of the construction the first step of the experiment is determined.

The following examples will show that the $n-1$ bound obtained in the theorem cannot be lowered. For each $n \geq 3$, define the machine D_n in accordance with the following table:

Machine D_n

Present State		
Previous State	Previous Input	
	0	1
q_1	q_2	q_2
q_2	q_3	q_1
...	...	...
q_i	q_{i+1}	q_{i-1}
...	...	...
q_{n-1}	q_n	q_{n-2}
q_n	q_n	q_{n-1}

Present State	Present Output
q_1	1
q_2	0
...	...
q_i	0
...	...
q_{n-1}	0
q_n	0

Then D_n is an $(n, 2, 2)$ machine such that any two of its states are distinguishable, but the shortest experiment which can distinguish q_n from q_{n-1} has length $n-1$.

For the case $n = 4$, D_n is represented by the following transition diagram:

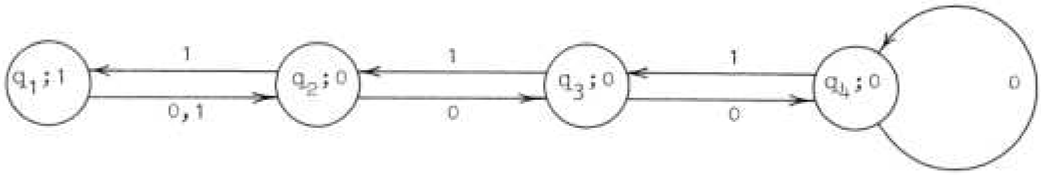

FIGURE 6. Transition Diagram of Machine D_4

Theorem 7: If S and T are (n, m, p) machines, such that some state q_i of S can be distinguished from state q_j of T, then this experiment can be of length $2n-1$.

First define the machine $S + T$, the direct sum of S and T. The table defining it will contain all of the entries and arguments of the table for S, plus entries and arguments obtained from those of the table for T by replacing q_i by q_{i+h}, for all i. This direct sum $S + T$ contains as submachines an isomorphic copy of S and one of T, but it is of course not strongly connected. Its transition diagram consists of the combined (but not connected) diagrams for S and T, with the names of the states of T changed to avoid ambiguity. Physically, the direct sum $S + T$ can be interpreted as a black box which has either the behavior of S or that of T, with no way of changing it between the two kinds of behavior. $S + T$ is a $(2n, m, p)$ machine such that certain pairs of its states are distinguishable, and hence by the methods used in proving Theorem 6, they can be shown to be distinguishable by an experiment of length $2n-1$. The experiment distinguishing any two states of $S + T$ also obviously distinguishes between the corresponding states of S and T.

The following examples will show that the $2n-1$ bound obtained in this theorem cannot be lowered. For each $n \geq 3$, define the machine E_n in accordance with the following table:

Machine E_n

Present State		
Previous State	Previous Input	
	0	1
q_1	q_2	q_2
q_2	q_3	q_1
...	...	...
q_i	q_{i+1}	q_{i-1}
...	...	...
q_{n-1}	q_n	q_{n-2}
q_n	q_{n-1}	q_n

Present State	Present Output
q_1	1
q_2	0
...	...
q_i	0
...	...
q_{n-1}	0
q_n	0

It can easily be verified that the shortest experiment which distinguishes q_1 of D_n from q_1 of E_n has length $2n-1$. For the case $n = 4$, the transition diagram of E_n is shown below:

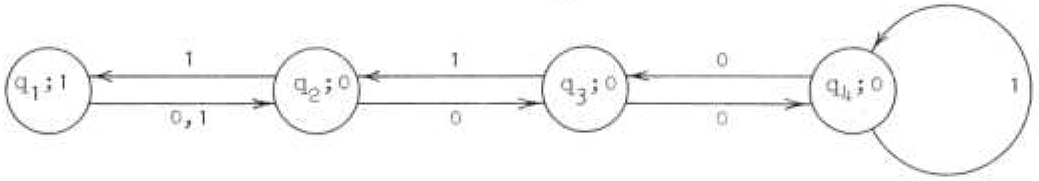

FIGURE 7. Transition Diagram of Machine E_4

Theorem 8: Given any (n, m, p) machine S such that any two of its states can be distinguished, there exists an experiment of length $n(n-1)/2$ which can determine the state of S at the end of the experiment.

This experiment will be constructed as it is being performed (since this will in general be a branching experiment, a complete formalization of this construction would involve defining a specific machine which could perform this experiment). As in the proof of Theorem 3, after each part of the experiment is performed there is a corresponding set of states which the machine could be in at the end of this experiment, i.e., which are compatible with all the outputs the machine has given during the experiment. Giving any one of certain sequences to the machine will reduce the number of elements in this set of states. Choose one of the shortest sequences having this property, and perform it as the next part of the experiment. Repeat this process until the set of possible states has just one element, i.e., the state of the machine is known.

It will be proved by induction on k that when the set of possible states of S has been reduced until it has $n-k$ members, at most $k(k+1)/2$ units of time will have elapsed. This is obvious for $k = 1$. For any $k < n$, let G_{k-1} be this set having at most $n-k+1$ members. Also the partition P_k, as constructed in the proof of Theorem 6, partitions the set of states of S into at least $k+1$ classes. Then G_{k-1} must have members belonging to at least two different classes of P_k (otherwise one class of P_k has at least $n-k+1$ members, and the other k have at least k members, so their union, the set of states of S, must have at least $n+1$ members). Consider such a pair of states belonging to different classes of P_k. An experiment distinguishing them has length k, and performing this experiment at this point will eliminate one or the other of the pair of states these will be transformed into by this experiment from the set of possible states of S. Hence by the fact that the shortest sequence having this property will be used in the construction, at most k more steps are required to reduce the set until it has $n-k$ members. Since by inductive hypothesis only at most $(k-1)k/2$ units of time had been used before this reduction, at most k more brings the total to at most $k(k+1)/2$. To complete the proof, let $k = n - 1$.

The following examples will show that the $n(n-1)/2$ bound obtained in this theorem is within a multiplicative constant of the best possible bound. For each $j > 3$, define the machine F_j in accordance with the following table:

Machine F_j

Present State		
Previous State	Previous Input	
	0	1
q_1	q_{j+2}	q_{j+1}
q_2	q_3	q_2
...	...	...
q_j	q_{j+1}	q_j
q_{j+1}	q_{j+2}	q_{j+1}
q_{j+2}	q_{j+3}	q_2
...	...	...
q_{2j-1}	q_{2j}	q_{j-1}
q_{2j}	q_{2j+1}	q_j
q_{2j+1}	q_2	q_1

Present State	Present Output
q_1	1
q_2	0
...	...
q_j	0
q_{j+1}	0
q_{j+2}	0
...	...
q_{2j-1}	0
q_{2j}	0
q_{2j+1}	0

Then F_j has $n = 2j+1$ states, is strongly connected, and any two of its states are distinguishable. It can be shown that the shortest experiment which can determine the final state of S consists of the sequence of length j^2+j-2 having a "0" in all positions except the first and those positions divisible by $j+2$, in which it has a "1". For the case $j = 3$, the transition diagram of F_j is shown below:

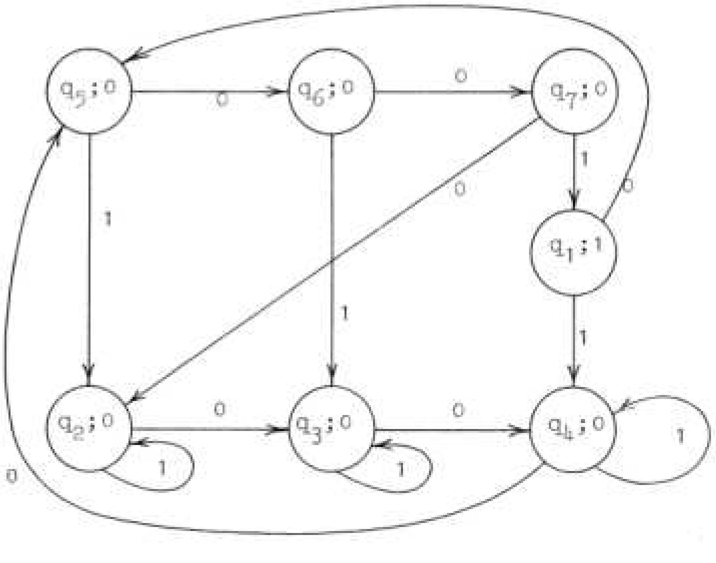

FIGURE 8. Transition Diagram of Machine F_3

Theorem 9: If $R_{n,m,p}$ is the class of all strongly connected (n, m, p) machines in reduced form, then there exists a simple experiment of length at most $n^{nm+2}p^n/n!$ which, when performed on a copy of any member S of $R_{n,m,p}$ will suffice to distinguish S from all other members of $R_{n,m,p}$

If $n = 1$, the result is obvious, so the proof will be concerned with the case $n \geq 2$. $R_{n,m,p}$ will be considered to have no two of its members isomorphic; that is, it will consist of just one of every essentially different strongly connected (n, m, p) machine in reduced form. Then define the machine Σ, the direct sum (as in the proof of Theorem 7) of all of the members of $R_{n,m,p}$. Apply to Σ the sort of experiment defined in the proof of Theorem 8, reducing the set of possible states it could be in until it has only one member. Then this identifies the machine S, up to an isomorphism.

To determine the length of this experiment, the first step is to note how many members $R_{n,m,p}$ has. Since there are exactly $n^{nm}p^n$ different (n,m,p) machines, the following correspondence between every member of $R_{n,m,p}$ and $n!$ different (n, m, p) machines will show $R_{n,m,p}$ has at most $n^{nm}p^n/n!$ members. In the case of any member T of $R_{n,m,p}$ having exactly n states, the correspondence can be direct with all (n, m, p) machines obtained from T by all $n!$ permutations of the names of the n states, since any two machines obtained from distinct permutations must be distinct. But if T is a member of $R_{n,m,p}$ having k states, with $k < n$, define the (n, m, p) machine T^* whose transitions and outputs agree with T for all q_i with $i \leq k$, but for all q_i with $k < i < n$, let the output be 0 and let all inputs cause a transition into state q_{i+1}, and for $i = n$ let the output be 0 and all inputs cause a transition into state q_1. Then the correspondence can be defined between T and the $n!$ different (n, m, p) machines obtained by permuting the names of the n states of T^*. Then since no two (n, m, p) machines have been made to correspond to the same member of $R_{n,m,p}$, $R_{n,m,p}$ has at most $n^{nm}p^n/n!$ members.

Then, proceeding as in the proof of Theorem 8, we can estimate how many steps must be necessary to cut down the number of possible states of Σ. It will be convenient to consider the subsets of states of Σ obtained from each of the original machines. By Theorem 8, at most $n(n-1)/2$ steps are required to eliminate all but one of the members of any such set. But by Theorem 6, this last state can be eliminated (unless Σ actually is in this state) in at most $2n-1$ steps. But $n(n-1)/2 + 2n-1 \leq n^2$ for $n \geq 2$, so each of the $n^{nm}p^n/n!$ subsets require at most n^2 steps.

It seems probable that the $n^{nm+2}p^n/n!$ estimate of this theorem could be improved considerably, since in the early parts of the experiment many states of Σ can be eliminated simultaneously. But it can be seen that the bound cannot be lowered below m^{n-1} by considering the following abstract model of a combination lock. For each $n,m \geq 2$, define a basic machine $G_{n,m}$ as follows:

Machine $G_{n,m}$

	Present State			
Previous	Previous Input			
State	S_1	S_2	...	S_n
q_1	q_1	q_1	...	q_1
q_2	q_1	q_1	...	q_1
...	...	...	...	...
q_n	q_1	q_1	...	q_1

Present State	Present Output
q_1	0
q_2	0
...	...
q_{n-1}	0
q_n	1

Then a combination lock will be defined as an $(n, m, 2)$ machine whose tables are obtained from those of $G_{n,m}$ by replacing, for each i with $1 \leq i \leq n-1$, exactly one of the q_1 entries in the ith row of the left-hand table above with a q_{i+1} entry.

The only way to make this give a 1 output is putting it into state q_n and this will be said to be unlocking the combination lock. If the combination lock is originally in state q_1, it can be unlocked only by giving it exactly the proper input sequence for the last $n-1$ steps before unlocking it. This input sequence is, of course, called the combination of the lock. The machine H is an example of a combination lock having the combination 0,1,0:

Machine H

	Present State	
Previous	Previous Input	
State	0	1
q_1	q_2	q_1
q_2	q_1	q_3
q_3	q_4	q_1
q_4	q_1	q_1

Present State	Present Input
q_1	0
q_2	0
q_3	0
q_4	1

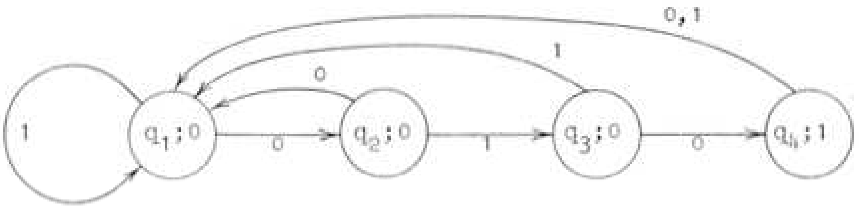

FIGURE 9. Transition Diagram of Machine H

For each n, m there are exactly m^{n-1} different combination locks. Suppose you are given some unknown combination lock, initially in state q_1, and are required to identify which one is present by an experiment which is as short as possible. Since in the first few steps in the experiment the lock cannot open, and at any later step in the experiment at most one combination lock can open, this experiment requires more than m^{n-1} steps.

FURTHER PROBLEMS

There are many further problems connected with this theory of sequential machines from the experimental point of view, which the author has not yet been able to solve.

One problem would be to find classes of machines more general than the strongly connected machines about which reasonable theorems can be proved. It should be pointed out that it was convenient to use direct sum machines (which are certainly not strongly connected) in two of the proofs. Infernal machines and the ordinary household electrical fuse provide important examples of machines which are not strongly connected.

Other problems which immediately suggest themselves are to improve the bounds given by Theorems 8 and 9. The author would like to conjecture, in this connection, that the best bound in Theorem 9 will be independent of p.

Still another problem of interest is the length of an experiment required to tell whether a given copy of an unknown strongly connected (n, m, p) machine S is indistinguishable from a known (n, m, p) machine T. This problem is akin to that faced by a maintenance man in checking whether a given machine is out of order. He knows what the machine is supposed to do, and he wishes to find out whether or not it does do this. If not, it is assumed that the machine is still a finite-state machine, differing in some subtle way from the supposed machine. A bound n on the number of states of the machine is helpful in view of Theorem 2, and is presumably derivable from the known number of relays or other components of which the machine is made. Theorem 9 does give a bound on the length of the experiment, although it seems fantastically large. A more reasonable experiment might be one which required the machine to undergo every transition only a few times.

Still other problems are suggested by permitting the inputs and the outputs of the machines to be k-tuples of symbols rather than single ones. The experimenters allowed in multiple experiments (see Figure 2) are already of this type, and many devices built out of relays or vacuum tubes have k-tuples of binary digits as their inputs or outputs. Such machines can be combined more freely than single input and output machines to make larger machines. Certain inputs of each machine are connected to the outputs of others, and other inputs and outputs of the individual machines are used as the inputs and outputs of the composite machine. If the k components of such a composite machine have n_1, n_2, ... and n_k states each, the composite machine has

$$\prod_{i=1}^{k} n_i$$

states, namely the k-tuples of states of the component machines. Such composite machines are of particular interest if all or most of these states are distinguishable. Many problems exist in relation to the inverse question of decomposition into such components. Given a machine with n states, under what conditions can it be represented as a combination of two machines having n_1 and n_2 states, such that $n_1 n_2 = n$? Under what conditions is the decomposition unique?

One way of describing what engineers do in designing actual automata is to say that they start with an overall description of a machine and break it down successively into smaller and smaller machines, until the individual relays or vacuum tubes are ultimately reached. The efficiency of such a method might be determined by a theoretical investigation on such decompositions. This might also throw light on the validity with which the psychiatrists can hope to subdivide the mind into ego, superego, id, etc.

Acknowledgements The somewhat overlapping work of Dr. D. A. Huffman has already been mentioned, and should be acknowledged as an entirely independent contribution. The writer would particularly like to acknowledge his indebtedness to Dr. C. E. Shannon for several suggestions which led to this work.