MACHINES CANNOT THINK

by

Robert George Gell

B.Sc., University of British Columbia,1962

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
M.A.
in the Department
of
PHILOSOPHY

We accept this thesis as conforming to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA
April, 1966

Department of Philosophy

The University of British Columbia
Vancouver 8, Canada

Date, 29 April 1966

ABSTRACT

This paper is a critical essay on the question "Can machines think?", with particular attention paid to the articles appearing in an anthology Minds and Machines, A. R. Anderson editor. The general conclusion of this paper is that those arguments which have been advanced to show that machines can think, are inconclusive.

I begin by examining rather closely a paper by Hilary Putnam called "Minds and Machines" in which he argues that the traditional mind-body problem can arise with a complex cybernetic machine. My argument against Putnam's is that either there are no problems with computers which are analogous to the ones raised by mental states, or where there are problems with machines, these problems do not have at bottom the same difficulties that human experiences raises.

I then continue by showing that a cybernetic machine is an instantiation of a formal system. This leads to a discussion of the relationship between formality and predictability in which I try to show that some types of machine are in principle predictable. In the next section I attempt to prove that any discussion of outward signs of imitative behavior presupposes that some linguistic theory, such as a type reduction, has been substantiated. The force of this argument is that such a theory has not in fact been substantiated. I offer some general theory about the complexity of concept-property relations.

Finally I give a demonstration that no test or set of tests can be found that will be logically sufficient for the ascription of the concept "capable of thought." If this is successful, then I have shown that no test can be found, which when a machine is built to pass it, is logically adequate for saying that that machine can think. This argument is offered as further criticism of the Imitation Game which A. M. Turing proposed as an adequate test for thinking subjects. Besides the specific conclusion that insufficient evidence has been offered to say that machines can think, this paper offers a more general conclusion that most standard problems have at bottom a linguistic difficulty. However this general conclusion is a broad speculative one to which the work in this paper is only a small exemplification and as such reflects mainly the further ambitions of the author.

ACKNOWLEDGEMENT

I wish to thank my two typists, Sue Reeves and April Toupin, and also Steve Porche who proof-read the manuscript and offered some valuable suggestions. Unfortunately I cannot acknowledge the help or encouragement of any of the members of the department in the preparation of this work.

I should like to dedicate this thesis to Lois, whose encouragement, though seldom acknowledged, was everpresent.

TABLE OF CONTENTS

Section Page

SECTION I

INTRODUCTION

This is a paper on the question "Can machines think?" and its general conclusion is negative. It is difficult to give an exact characterization of the problems that philosophers are interested in when they discuss this question. However it would be fairly safe to say that the problems are those posed by the recent advances in digital and analogue computers. These machines have been built to perform a great variety of human tasks and the question naturally arises as to whether or not we must say of some 'super' computer that it thinks. In this respect, of course, it is of interest to consider the definition of a mechanical computer to see if there are any limitations serious enough to justify us in withholding the designation, 'capable of thought'. Before we can decide whether or not a machine thinks, a great number of secondary problems must be tackled and these problems are of wide general philosophic interest. Furthermore the philosophic importance of recent developments in mathematics and physics must also be assessed. So potentially the problem "Can machines think?" could lead us into very general philosophic speculation. However, an article by A. M. Turing[1] in 1950 sparked a whole series of papers in the philosophic journals, some of which were collected by A. R. Anderson, in an anthology called Minds and Machines.[2] This paper is a criticism of the main arguments presented by those who feel that machines can think with particular attention given to those articles in Minds and Machines.

There are several conclusions arrived at in this paper.

The argument in the second section attempts to show that there is no serious analogy between men and machines. That is to say, no serious analogy in the sense that those problems which are raised because of the uniqueness of human experience, are not raised with very complicated computers. In the third section I show that a Turing machine is formal, and as such is, in the important sense, predictable. The fourth section is an attack upon the possibility of building a computer to imitate human behaviour. The argument is, that until certain things are shown about our behavioural concepts, then the problem of imitation cannot arise. Of course the force of this argument is that these things have not been shown. Finally in the fifth section I try to show that no test can ever be constructed which will logically be adequate for the application of the concept, 'capable of thinking'. This argument is meant to undercut the long debate which has gone on criticizing Turing's "Imitation Game",[3] which was proposed as a test for thinking.

Most of the arguments of this paper are an exemplification of a general philosophic approach. This approach is one in which attention is focused on the concepts that we use. By doing this attention is drawn to the complexity of these concepts, particularly in their logical structure. It is argued that too little attention is given to the complexity of language, particularly with respect to our behavioural concepts. At times it is argued that until some problems about the nature of concepts are answered, then no decision about the possibility of constructing robots can be made. So in a sense, I think that the general nature of the thesis of this paper can be said to be linguistic. There is also a larger thesis behind this paper, but upon which none of the arguments depend. This is the idea that most questions of philosophic importance can be put in the form of a problem about the logic of concepts. If this is so, and a systematic way can be found for discovering the logic of concepts, then the main problems of philosophy can be solved within a science of language. This paper does not attempt to establish this thesis but rather is meant to be in some small way an exemplification of it. Thus the arguments of this paper try to show that the problems connected with thinking machines can all be given a linguistic interpretation; although no attempt is made to give a method for discovering the logic of concepts.

I should mention, finally, that this paper does not reiterate in any detail, the arguments which have already been made in the many papers on this subject. In fact it is assumed that the reader is familiar with most of the arguments and in particular that the reader is very familiar with some specific articles. In some places this paper is an extension of some very thorough work by other philosophers. But in general the criticisms of this paper are very broad and are intended to undercut many of the standard ideas connected with the problem "Can machines think?"

SECTION II

THE ANALOGY BETWEEN MEN AND MACHINES

Hilary Putnam in his paper "Minds and Machines" tries to draw an analogy between the various states of a complex cybernetic machine (called a Turing machine) an d the corresponding states of a human being. He maintains that a machine has logical and structural states, just as a human has mental and physical states, and also that those arguments which support the identity or nonidentity of mental and physical states also show that the same thing about logical and physical states. As we all know, a machine is capable of a complete mechanistic (causal) explanation and has no hidden or otherwise mysterious parts. Thus if Putnam can sustain his analogy between men and machines, he thinks that this will go some way (he does not think it would be conclusive) in substantiating a mechanistic thesis. It is my contention in this section that Putnam fails to find the analogy that he is looking for.

Putnam's thesis rests on two main claims. He tries to show that the proposition "I am in state A if and only if flip flop 36 is on" is, from the machine's point of view, synthetic, or what is taken to be the same thing, at least empirically verifiable. This will make it analogous to the proposition "I am in pain if and only if C-fibres are stimulated" and will depend upon there being different methods of verification of state A and flip flop 36 being on. His other claim is that the logical-structural distinction is analogous to the mind-body one in that there can be a logical description of the machine's computations just as there is a mental description of human activity. I hope to show that even from the point of view of the machine, the above proposition is not synthetic, and also that the logical-structural distinction is not analogous to the mind-body distinction.

Putnam considers a Turing machine 'T' which can be in a number of states, one of which is named A. As he says, "a Turing machine is a device with a finite number of internal configurations, each of which involves the machine's being in one of a finite number of states,..." I presume that any particular state of T is defined as a unique combination of certain circuits being activated, certain circuit breakers being open, and certain vacuum tubes operating and further that other circuits are dead, other circuit breakers are closed and other vacuum tubes are not operating. It may be the case however that the condition of some components of the machine are irrevelant in the determination of some state, (say) state A. For the discussion in this section , let us define state A as that state of a Turing machine in which flip flop 36 is on and all the other circuits are irrevelant. This last clause, "and all the other circuits are irrelevant" can be expanded into a finite list. Instead of specifying whether the other components of the machine should be closed or non-operational, we can say if some circuit which is irrelevant that it can be either open or closed, either operational or non-operational. In this way we can expand the definition of state A into a finite list, such as, flip flop 36 is on, flip flop 1 is either on or off, flip flop 2 is either on or off, etc. As I said, this description is finite because there are a finite number of components in any machine. We can now generalize our description of what a state is by saying that any state of a Turing machine is equivalent to a list of the various components of the machine stating either that they are on, off, or either on or off; active, inactive or either; active or inactive; etc. Thus any particular state could be pictorially represented by a plan of the machine showing the conditions of the various circuits, circuit breakers, tubes, magnetic fields, relays, etc.[6]

We could build into T a sub-machine (sub-T) which could check the condition of the various components of T and which would print-out (say) onto the input tape of T, the results that it obtained. If we wish to check for the various states that T is in, it will simplify our job considerably if we determine what are the sufficient features of each particular state which differentiate it from all the other states of T. Then we could speak of the sufficient conditions for any particular state. There will be many of the configurations of state B which are different from C or D but not from E or F. But that configurations of the various components of T which is sufficient to differentiate some state from all the others will be called the sufficient conditions of that state.

Now that we have this machine, we can ask it to verify the statement "I am in state A when and only when flip flop 36 is on." To give a plausible situation for this to arise, imagine that we have just built T and theoretically the position of flip flop 36 should be the sufficient condition for state A. We ask the machine to check (or as Putnam considers, the machine itself considers checking) the above statement. The method would be theoretically simple. The machine enters state A and sub-T reports the condition of flip flop 36. The machine then enters every other state (which is a finite number) and compares the reports of sub-T on flip flop 36 to the first report. If the subsequent reports are all different than the first one, the proposition is true. There is however a vast practical problem of getting the machine to go through every other state, and making sure that none are missed. However, this aside, the statement seems open to an empirical solution, making it synthetic.

Putnam wants to say[7] that if some bright person raised the question of the identity of state A and flip flop 36 being on, the same objections could be raised against identity in the machine case as are raised in the case of the identity of being in pain and C-fibres being stimulated. In the mind-body case it is argued that since there are different ways of knowing about the states to be identified, the two states could not be identical. These same considerations hold in the machine case. The way that T determines the state of flip flop 36, is from the reports of sub-T, and the way that it determines what state it is in, is from the original input order to enter the state. So there are two different ways of knowing about the two states. Thus state A is not identical with flip flop 36 being on. At this point Putnam leaves the reader with the choice of saying either there is a 'mind-body' (or logical-physical state) problem with machines or else the human mind-body problem is merely linguistic. Before we take Putnam's choice, let us go back and see whether or not the considerations are actually parallel.

I gave an example earlier in which the statement "I am in state A if and only if flip flop 36 is closed" was synthetic. But the example I gave to illustrate that, was the case of checking the operation of some machine which had just been constructed. It is a normal assumption in the discussion of machines that we are only considering 'theoretical' machines; i.e., those that never have mechanical failures. I assume that Putnam is talking about the same machines that Turing,[8] Church,[9] and Davis[10] were, and these were theoretical machines. If T is a theoretical machine then, the case I gave to illustrate the synthetic nature of the statement could not arise. By dealing with theoretical machines we eliminate the possibility of malfunction in the machine, so the problem of seeing whether or not the machine functions as designed cannot arise.

But perhaps there is a further sense in which the statement is synthetic. Isn't it an empirical question as to whether or not the position of flip flop 36 is the sufficient condition of state A, i.e. is the position of flip flop 36 the feature of the internal condition of the machine which makes the state A different from all other states? But this question is not the original question but rather the one as to whether "I am in state A <u>if</u> flip flop 36 is closed." This is of course rather obvious because the necessary and sufficient conditions are a complete description of the circuits, circuit breakers, and tubes being on,off or either, in the proper configuration for state A. That these configurations are the proper ones is not an empirical or synthetic question but rather a question of naming or defining just which configuration would be state A. Since th statement "I am in state A if and only if flip flop 36 is closed" is ane about the necessary and sufficient conditions of state A, it is a matter only of the way the machine was set up; i.e., a matter of the initial stipulation.

That thedefinition of state A is a matter of initial stipulation though, does not prevent the question about the definition of state A being asked. The machine may consider, or some programmer unfamiliar with T may consider, the truth of' the proposal "I am in state A if and only if flip flop 36 is on". This will be a difficult, but not insoluble problem, but this alone will not show the proposition to be synethic. The initial assumptions of any system, or the orginal constructional correspondences of any machine, may be difficult to determine but this does not prevent them from being stipulations (or axioms or definitions). Thus the fact that there is quite a problem, which one may fail to solve, in ascertaining the initial stipulations of the various states of the machine, does not show that these correlations (namings, stipulations) are not analytic.

The main argument however is, that the ways of determining state A and the position of flip flop 36 are different, and thus it seems an entirely contingent matter whether or not the two things are identical. "For instance," Putnam says, "the machine might be in state A and its sense organs might report that flip flop 36 was <u>not</u> on." In which case the machine would have to decide whether to say the proposition was false or to attribute the discrepancy to observational error. This problem which Putnam poses for the machine could never arise with a Turing machine because we are assuming that the machine functions correctly for as long as we want it to. So there is no possibility of an observational error in a Turing machine, and if there was an 'observation' of flip flop 36 being off when the machine was in state A, then the only conclusion is that the given statement is false. But if the exact problem which Putnam raises can not arise, still we have the fact that there are two independent ways of verifying each part of the proposition, [illegible] The way the machine determines the position of flip flop 36 is from the input report of sub-T. But how does T determine which state it is in? The machine determines this from the initial input order which was given it (or even which it gave itself.) At no time does the machine directly observe that it is in state A as Putnam claims. The machine infers from the evidence of the input order to the actual internal configuration. Also the machine infers from the evidence of the input results of sub-T to the actual internal configuration. Thus in determining whether it is in state A or whether circuit breaker 36 is on, the machine makes an inference from evidence which is presented to it. Although the evidence is different, the method of verification is the same in both cases.

Furthermore, since we are dealing with theoretical machines, we assume that no mechanical failures occur and that there have been no mistakes in programming. So, for a Turing machine it is not possible that T be given an order and fail to execute it or that sub-T report incorrectly. Thus both the input order and the report of sub-T become definitional criterion for state A. Therefore, if the proposition "I am in state A" means that the machine has been given the order to enter state A, either by itself or some programmer, and since flip flop 36 being on is a necessary condition for state A, then the proposition "I am in state A if and only if flip flop 36 is on" is analytic for a Turing machine. On both accounts then the case of the machine is different from the human case. The proposition "I am in state A if and only if flip flop 36 is on" is analytic whereas the analogous proposition "I am in pain if and only if my C-fibres are stimulated" is synthetic. The ways in which the machine verifies both the state it is in and the condition of flip flop 36 are the same. Whereas, in the human case there is an in-principle difference between the ways of verifying that one is in pain and that one's C-fibres are stimulated. So Putnam has not built an analogous case with Turing machines.

Putnam then turns to showing that the question of whether a machine 'knows' what state it is in, is a degenerate question.[12] If he can show that it is degenerate in a way that the similar question about human knowledge of mental states is, this will add more evidence to the analogy between logical states of a machine and mental states of a human. So he compares the two questions "Does the machine 'ascertain' that it is in state A?" and "Does Jones 'know' that he is in pain" in order to show that questions about the method of attaining knowledge of internal machine states. He hopes to show that they are both degenerate, but I shall argue that the questions about machine methods are either not degenerate or if they are, not for the same reasons that questions of method are for mental states.

There is one obvious sense in which it can easily be said that the machine computed state A, and that is the case where the machine goes through a series of calculations which terminates in state A. But I take it that Putnam is interested in the case of whether or not a machine can be said to compute that it is in state A from state A alone. Before considering the question though, we must add one more feature to our machine T, by supposing that whenever the machine is in one particular state (say state A), it prints out the words "I am in state A". This can be done in two ways: either every time we give the machine an instruction to enter state A, we next give it the instruction to print out "I am in state A", or else we can have the machine so constructed that every time it enters state A it also prints out "I am in state A". The question may now arise "Does the machine 'ascertain' that it is in state A?" According to Putnam, 'ascertain' is synonymous with 'compute' or 'work out'; so the question can be rephrased as "Does the machine 'ascertain' (or compute or work out) that it is in state A?"[13] If we have a machine in which a further instruction is given it to print out "I am in state A", then the answer to the above question is yes, and the answer to the further query about how it ascertains or works it out is given by showing the programming required. In this particular case it is a matter of the insertion of a sub-routine (granted it is a short one of one instruction) after the instruction to enter state A. So if we have this type of machine, the question is not degenerate.

But if we have a machine that has built into it a programme such that every time it enters state A it prints out "I am in state A", then the printing out becomes part of the description, and thus a definitional condition of the machine being in state A. (Mechanical errors are theoretically eliminated.) If this is the case then it loses its analogy with the human situation of someone 'evincing' "I am in pain", for the verbal statement is not part of the description of pain and not a definitional condition of being in pain. The question about the machine ascertaining or computing that it is in state A becomes degenerate because the fact that the machine printed out "I am in state A" is a definitional criterion of the machine's being in state A. Putnam says that the difficulty of degeneracy has, in both cases the same cause: "namely, the difficulty is occasioned by the fact that the verbal report (I am in state A and I am in pain) issued directly from the state it reports..."[14] But the print out "I am in state A" is not a report, but a part of what is stipulated as being in state A; reports can be mistaken, but not definitional criterion. The question about the machine computing "I am in state A" from state A is a description because part of what is set up in this machine asstate A is a description of the print-out mechanism printing "I am in state A", and not as Putnam thinks because "I am in state A" issues directly from the machine's being in state A. However the statement "I am in pain", if it is degenerate, is not so for these reasons. In the human case, a person saying that they are in pain is not a necessary condition either for them knowing themselves that they are in pain nor for someone else knowing that they are in pain. The relation between the statement "I am in pain" and the pain is quite contingent, and it is this fact which gives rise, in the human situation, to the question of knowing about the pain in order to 'evince' "I am in pain". This analogous situation does not arise in a Turing machine. So the question of how a machine computes or works out what state it is in, is not usually degenerate, but when the question is, it is not degenerate for the reasons that questions of knowing pain (if those questions are actually degenerate) are.

To continue his analogy betwe̶en machines and humans, Putnam shows that there are two types of machine states, logical states and structural states, and that these are analogous totthe mental and physical states of human beings. As I mentioned earlier, any theoretical Turing machine is capable of being in a finite number of states, A, B, C, ..., and if the various programmes of this machine are already in memory, then the machine will change from one state to another according to its programming. But as Putnam says "a given 'Turing machine' is an abstract machine which may be physically realized in an almost infinite number of different ways,"[15] and, for any particular manufactured machine the physical condition of it mayvvary from one condition to another. Thus any actual machine may be in a number of physical or structural states and yet may still be in the same logical state. So for any particular machine it can be thought of or described as a finite number of logical states or as a number of structural states, and the functioning of the machine can be expressed either entirely in terms of logical states, or again, entirely in structural states. This is, according to Putnam, analogous to the human situation in which the functioning of the human can be explained in terms of mental occurrences (e.g.,Freudian explanation) or in terms of physiological changes (e.g., complete behavioural description).

In order to assess this analogy, let us backtract to the distinction between logicaleand structural states and consider briefly again just what are logical states. When we set up a Turing machine, we said that it could enter a finite number of states, A, B, C, ... etc. These states referred to something more or less explicit; namely the internal configuration of some hypothetical machine. These states of the machine, A, B, C, ... must be explicit, at least to the extent that we can see that we can build some machine that will enter these states. Thus if the particular state we are talking about is one in which the machine places the input data into memory space 4683, we must be able to show that a machine can be built which will fulfil this function and consequently be able to enter this state. This could be done by laying out on the drafting board the possible configurations of circuits, relays, and vacuum tubes such that any machine which was built from these plans would be able to enter this particular state. This requirement that the states of a Turing machine refer at least to one possible configuration of a machine, is absolutely essential. Otherwise we would beg the entire question. If we simply said that the machine could fulfil such-and-such function and we did not specify how this could be accomplished mechanically, then we would simply be saying that machines can do whatever humans can and I presume that it is just this question of whether machines can do everything humans can do that we are trying to answer. So unless we beg the question, we must be able to specify at least one mechanical configuration of a possible Turing machine for every state that we attribute to machine T. When we say that the internal configuration of state A must be specified, we do not mean that it must be explicitly laid out in every minute detail. For example, if in specifying state A we say that there must be a circuit joining the scanner to the memory input compartment, we do not specify the length of the circuit, nor the chemical composition of the wire, nor even for that matter that it must be a wire which carries the impulse from one to the other. In fact there is no limit to the various ways that such a circuit could be set up. (The circuit is specified by the function, (or purpose, or goal), and thus there are an unlimited number of actual mechanical ways of fulfilling the particular purpose. We could also have a messenger boy carry the message, but this would not be a mech anical solution. But we must show that there is at least one mechanical solution.)

On the other hand, for any actual machine there will be a complete physical description of the various circuits, relays, tubes, etc., specifying the actual physical make-up of the machine. But these specifications must include at least those specifications which were laid down for the theoretical state. That is, those conditions which we specified for the T machine to be in state A must be included in (or deductible from) the physical specifications of this actual machine, although these physical specifications will also describe many properties which were not included in the theoretical considerations of state A. Our initial specification of the properties of state A was abstract in the sense that it left open to the engineer building the machine many other properties to be specified before the machine could be built. But the computer's physical or structural description of state A will differ from the theoretical or logical description of state A only in that it describes more properties for the machine. Thus if we think of the structural description as designating a set of properties and conditions of T, the logical description will be a sub-set of these.

Now it is usually thought that the difference between mental states and physical states is one of a more serious nature than just that mental states have the same but fewer properties than physical states. It is generally thought that the test for determining physical properties are not applicable to the properties of mental states. Most of the philosophical speculation of the last few years has been an attempt to find some identity principle between the properties of our mental states and those properties which are objectively attributed to other people. Putnam doesn't even need an identity principle because there is only one type of property. He has failed to find two types of things between which we need to find some bridge or connection. From a complete physical description of a machine we can deduce the theoretical description, but until some identity principle is afforded by Putnam or someone else, we cannot deduce the mental description of a person from his physical condition. This identity principle which would bridge the gulf between mental and physical states may yet be found by philosophers, nevertheless, what is certainly true is that some principle is needed. In the case of a Turing machine there is no principle needed because Putnam has failed to show that there is a type difference between the properties of logical and physical states. Therefore the difference between a logical and physical description of a machine is not analogous to the difference between a mental and physical description of some person's pain (say). Thus I conclude that the logical-structural distinction with machines is not analogous to the mental-physical distinction in the human situation.

The conclusion of this section is not that there are no problems to be answered or distinctions to be made with complex Turing machines. The conclusion is rather that the problems raised or the questions asked by a Turing machine about itself are not problems for the same reasons that similar questions about humans are. The machine may ask itself questions of the same form as humans may, but the difficulty is not the same difficulty that a human has. Similarly, many distinctions can be drawn in dealing with complex machines, but these also, I conclude, are not the same distinctions which philosophers have noted in the human case. Thus the problems which a complex Turing machine might face are not the same as those that humans try to answer, and in this sense the analogy between men and machines fails.

SECTION III

FORMALITY AND PREDICTABILITY

In this section, I wish to show that a Turing machine is a concrete instantiation of a formal system, and as such, is predictable. My demonstration that Turing machines are formal is not unique but I feel that it is important that it should be shown rather explicitly. Many philoso phers have argued that if a Turing machine is formal then Godel's Incompleteness Theorem can help us to some interesting conclusions about machines. Some, such as Lucas[16] have argued that the Theorem refutes mechanism; others, such as Putnam[17] and Turing[18] have argued that the Theorem has no bearing on the interesting philosophic questions. I shall argue, on the other hand, only that Turing machines are formal and that in the important sense that philosophers have concerned themselves with, these machines are predictable.

Before entering the problem of showing any limitations of a Turing machine, we must demonstrate rather clearly that any Turing machine can be represented as a formal system. My demonstration of this is essentially the one used by Martin Davis in the first chapter of his book, Computability and Unsolvability.[19] As I explained in the first section , a machine can be in any one of a number of configurations, q_1, q_2, q_3,... up to some finite limit. A tape, divided into discreet units, is fed into the machine and in each unit there appears a letter of a language comprising a number of symbols, S_0, S_1, S_2,... up to some finite number. Furthermore the tape is finite, but can be as long as is needed. One of the essential functions of a Turing machine is that it is able, upon the receipt of a symbol, to change from one state (say) q_1 to another state q_2. Not only can a Turing machine change states but it can also change the symbol on the scanned unit or it can move the tape along so that the next unit is scanned. This possibility of changing can be represented by a quadruple, such as, $q_1S_1S_2q_2$. The machine that this is a quadruple of, will, if it is in state q_1,and is scanning symbol S_1, change to state q_2, and erase S_1 and put the symbol S_2 in the scanned tape unit. More generally, a quadruple stands for a machine built to carry out any instruction of the following form: when in state $q_x$and the symbol S_x is on the tape unit being scanned then change to state q_y ($x \geqslant y$ or $y \geqslant x$) and either change the symbol on the scanned tape unit to S_y or else scan the unit to the right or left. If a machine is capable of following out an instruction of that form, then it can be represented by a quadruple. It is important to notice that after the machine has carried out this instruction, it is in the original position again in that it is in some state with a scanned unit in front of it. Thus the machine is ready to carry out another instruction of the same form. However, if there is no such instruction built into the machine, then when it reaches that state and symbol, the machine will stop. Thus any machine which goes through a process or series of changes from one position to another can be represented by a series of quadruples. Since the number of states and symbols is finite, the number of quadruples will also be finite. Therefore all the possible movements of the machine can be described by a series of quadruples, so that this series actually defines the machine"s possibilities.

Any particular Turing machine can be represented, then, by a series of quadruples. But as I said, when the machine has finished one change it is in a position to carry out another. This continuous change of the machine is represented by a series of deductions. If we take the tape to be given for any particular machine, then by knowing which unit the machine will scan first and the state that the machine is in when it begins, we can deduce, using the list of quadruples of that machine, the various steps that the machine will go through to arrive at the answer. So considering the q's and S's as primitive words, and the original tape as initial axioms, and the quadruples as rules of inference, we have constructed an axiomatic system which with the addition of a few more stipulations can be made quite formal. And this system represents, in symbolic terms the various changes that a Turing machine would go through in any actual problem. I shall in what follows, state this fact rather briefly by saying that a machine is a concrete instantiation of a formal system.[21] Finally, any theorems which apply to formal systems, as formal systems, will also apply to Turing machines.

If we consider a computer as a discrete state machine whose motion follows some formal system, then it seems that whatever the machine does is predictable. If we know the initial state of the computer and we know its complete list of quadruples then we can predict what the machine will do once we see its tape. However, does it follow from the fact that a machine is formal that it is predictable, and further, if the machine is not predictable does this show that it is not formal? Now there are several reasons to suggest that a computer is not predictable. One reason may be that we do not have enough knowledge of the machine. For example, we cannot predict (in general) when a complicated piece of machinery will break down because we don't know enough about the manufacture or structural composition of the various parts. But we attribute the inability to predict simply to our lack of knowledge which we feel that we could get, given enough time and laboratory space. That is, we hold that for these reasons machines are not 'in principle' unpredictable. However, there are other reasons for the unpredictability of computers which stem from our inability to get knowledge. But this inability is not a practical matter but a theoretical one. I take it that the implication of Heizenberg's Uncertainty principle is that measurements below fixed amounts are not possible, for the more accurately we measure the position of a particle the more inaccurate will be our measurement of its momentum. So much so that if we ever did measure the position of a particle completely accurately then we would necessarily have made an infinite error in its momentum. Thus, considering measurements of the utmost accuracy, we must, in principle, have a finite magnitude of error, and we are unable to predict anything to an accuracy greater than the accuracy of the accumulated errors. However as I said, we are dealing with measurements of great accuracy and of course we will be measuring sub-atomic structures. For if we want to make a measurement of something to the greatest accuracy we will have to consider the object as a collection of sub-atomic particles. But if we consider the object or machine as a macroscopic unit, then using macroscopic measuring devices, we can, within experimental error, measure, test, and predict the movements of the mechanism. So if we spent a great deal of time testing the various parts of some machines, the above reasons would not be sufficient to show that any machine is in principle unpredictable in macroscopic units.

It is generally contended, however, that computers which contain randomizing devices are in principle unpredictable. I want to examine two types of randomizers, (a) a counter of the number of radium atoms to have disintegrated in the half-minute previous and (b) the decimal expansion of π . I take the counter as an example of a device which we can never, regardless of how much knowledge we had, predict, i.e. the number which the counter has on it at any moment is in principle unpredictable. The reason for our inability to predict may be due to the variations which affect the disintegration of radium atoms being of such a small magnitude that the Uncertainty Principle limits our investigation. (This would only show that we cannot investigate the laws governing disintegration although there may be some.) But granting that there are in the world counters which are unpredictable in the strong sense that no increase in knowledge will ever avail in predicting them, what can we say about computers which contain these devices?

Presumably, a computer with a random device will work as follows, the machine is given the instruction to look at the tape unit totthe right and there is no symbol on that unit. The symbol is not written on the unit until the tape is in the scanner and then the symbol which is written on the unit is determined by the random device. In this way no one could predict how the machine would operate after this instruction because we could not, in principle, know what symbol would be on the tape until the machine actually did scan the unit. However this example is just another case of adding more information to the machine during its calculations. We can certainly build machines that will do some calculations and then come to a halt until more information is given to it. This would be the case where the machine works out the initial tape input, and when it stops we alter the tape, which is just the same as giving it a new tape. Then the machine will work again this problem. We can make this more sophisticated by having the machine itself add more information to the tape at certain stages of its calculations. And the case of having a randomizing device in the machine is an example of adding more information, but the information can not be predicted.

When we originally thought of the problem of predicting a computer, we were thinking of a machine which was given some calculations to do. In terms of the machines formal system, the case of predictability arose where we had a finite list of quadruples and a given series of tape expressions. Then it was asked whether or not the machine's movements could be predicted. This is all quite analogous to the human situation where we give someone a problem and then try to figure out what their behaviour will be. But the original problem was not one of trying to predict how a machine would react when given more information later in the problem, information which we could not get ourselves. No one would think that you had shown a machine to be unpredictable if you proved that we cannot figure out in advance how the machine would react when unknown information was fed into it. When we ask whether or not machines are predictable we are asking whether or not, given a machine and the information fed into it, we can predict the subsequent movements of the machine.

The randomizing device feeds information into the machine from within the machine. But I do not think that this changes the case at all. The tape that the machine scans is changed and that creates a new axiomatic beginning for the machine. The fact that the source of the information is some device within the physical bounds of the machine does not make the case different than the one where more information is fed in from outside. It may be thought, however, that I am prejudicing the case by making the randomizing device peripheral to the actual machine, and that actually the device can be built into the 'essential' workings of the machine. I myself cannot see how this randomizing effect could be expressed in terms of quadruples and tape expressions except in a way similar to the one suggested above. If we build the device into the essential workings of the machine, then we would not have a computer but rather just a super-randomizer. The purpose of a randomizer is to supply random numbers when the machine requires that type of information, viz. random numbers. Therefore, a computer with a randomizer is still quite predictable as far as its movements are concerned during a problem. It is not predicable, however, if during the problem more unknown information is fed into the machine, but then no one ever thought that a machine was predictable under those conditions.

If the type of randomizer is one that selects numbers successively from the decimal expansion of π, then the computer is completely predictable. If we build the machine so that each time it receives an instruction to 'search', it selects the next number successively in the expansion, then the numbers which the computer selects will be random. However if we know how many past searches the machine has done, and we know where in the expansion the computer started, then we can calculate the next number and we will know which alternative the machine will follow. Thus there are machines with randomizers which are together completely predictable. We can conclude, therefore, from the discussion of the two types of randomizers, that computers with these devices in them are still predictable in the strong sense. Furthermore the formality of the machine is not upset, because we can easily allow for a change in the input tape, which we said was comparable to the axioms of a formal system. Altering the axioms of a system does not destroy the formality of the system, it just makes a new system that has different theorems.

SECTION IV

WHAT IS BEHAVIOUR?

In his article "The Mechanical Concept of Mind",[22] Michael Scriven presents the following argument:

> "... the outward signs (including speech) are not infallible indications of consciousness. It is therefore quite certain that they are not, ... the same thing as consciousness."[23]

This argument is meant to show that consciousness cannot be reduced to outward signs or observable behaviour. Scriven seems to have in mind a distinction between the behavioural and the non-behavioural aspects of man. When he talks about two distinct things, outward signs and consciousness, Scriven seems to be distinguishing between outward observable behaviour and something else which is inner and unobservable. In order to assess this argument which I have quoted or any others like it, we must make clearer this distinction between outward signs and consciousness. In particular, it might be asked just what are the outward signs? What are the behavioural aspects of man? More generally, this is just the qustion "What is behaviour?"

When philosophers talk about the possibility of there being mechanical robots around, it seems that they are also using the idea of a robot to mark the distinction between the behavioural aspects of human experience and the non-behavioural aspects. The robot is considered to be able to behave exactly like a person, even, with some writers, to the point of being bahaviourally indistinguishable from other people; so that whatever else a man has besides behaviour, that's what makes him different from a robot. No one ever considers actually building a robot and philosophers are not interested in some supposed future problem of distinguishing actual people from their mechanical robot slaves! When we conceive of mechanical robots, we are just using a conceptual device to mark the distinction between those things which have just behaviour and those which have something else besides. Again, however, before we can consider using this conceptual device of mechanical robots, it is important to determine just exactly what is to be considered as behaviour.

It is generally thought that if we could build a robot to imitate any human behaviour, that we would not be able to differentiate the robot from other people as far as its behaviour was concerned. However, even if we grant that a machine could be built to imitate any human behaviour, this would not mean that it was indistinguishable from a human. The fact that we can build a robot to imitate any piece of human behaviour does not prove that we can build a robot to behave the same as a human. We do not usually equate 'acting like' someone else and 'imitating' them. Take the case where X is said to be imitating Y. If we could show that X was unaware of what Y was doing, then we could not say that X was imitating Y. Furthermore if we are correct in saying that X is imitating Y, then we could correctly attribute some intention to X; namely, the intention to imitate Y. Whereas when we say that so-and-so is acting like another person we are implying only coincidence. Confusing 'acting like' and 'imitating' is tantamount to reducing coincidental behaviour to conventional behaviour, like confusing similar and typical. There is certainly a difference between on the one hand, two people having similar enough characteristics to be indistinguishable and, on the other hand, people having certain characteristics the same but not having some others. Imitating is a case of having some characteristics the same as whoever is being imitated but not having some further characteristics. Acting like or being alike is a matter of doing similar sorts of things, things which are comparable enough to be called the same. So if we allow that imitation of any piece of behaviour is possible we can not move immediately to the conclusion that robots and humans are indistinguishable. Thus if we allow that robots can be built that imitate human behaviour, it by no means follows that they are indistinguishable, even behaviourally, from humans.

This claim, that if we allow that imitation is possible does not prove that men and robots are indistinguishable, is quite compatible with the evident fact that during a performance an actor may be indistinguishable from (say) someone who is really mad. For to say that an actor is indistinguishable during a performance is to admit (tacitly) that there is a definite limit to the similarities between actors and madmen. But to admit that there are limits is to acknowledge that actors and madmen are readily distinguishable in a larger context. However, there may be cases of imitation which are done so well that one may doubt whether there are any characteristics which the imitator has failed to duplicate; a sort of perfect imitation. But I find this case generally inconceivable, since imitating presupposes a (particular) second-order intentionality on the part of the actor which the person imitated doesnot have. Unless one held that intentionality was entirely non-behavioural, i.e., had no behavioural manifestations, then I cannot conceive of a case of perfect imitation. However it is certainly the case that if we allow that robots can be built to imitate any piece of human behaviour, we can not conclude from this that they would be indistinguishable, even as far as their behaviour itself is concerned, from humans.

However, let me try to make clearer the distinction that I drew above between the problems of similarity and exemplification. In problems of similarity we are trying, for example, to determine whether some particular piece of behaviour can be called a smile. We are troubled because we donot have any clear test for determining what constitutes smiling. Or, we may be in doubt about how successful one must be in some proposed test in order to be said to have smiled. This is the problem of trying to find adequate tests for the application of some characteristics to given situations. By an adequate test, I mean one that is successful or positive when we say that the situation has the characteristic, and unsucessful or negative when the situation doesnot have the characteristic attributed to it. This means that the statement of the success of an adequate test is logically necessary and sufficient for the statement of the description of the characteristic to the given situation. Thus when we raise questions about adequacy, what is in doubt is the relationship between the characteristics attributed to some situation and the tests done on the situation.

However, in the other problem of finding typical examples, we may be in doubt as to whether two subjects have the same characteristics because the test we have will not apply to one of them. Or, if we can see that they both have some characteristic in common, we may try to find some other characteristics which one has and the other hasnot. This problem may arise, thinking now of robots, in which someone says they have produced an example of something claiming that their product has all the characteristics of the other things. This is the problem of determining the characteristics of any given situation which one may select to examine. Thus there are two distinct problems: that of determining the adequacy of tests and that ofddetermining the various characteristics of given subjects.

I do not think that these two problems are unrelated; in fact I shall argue that one presupposes that the other has been answered. It can readily be seen, I think, that in order to answer the question of whether or not some proposed subject is to be admitted to another class of objects as a typical example, we must have some way of determining the characteristics of the members of the group and also of the proposed subject. If the proposed example has all the characteristics of the members of the group (that are relevant to them being a group), then the example becomes a member. But this problem could not be tackled until we have some adequate way of deciding when two subjects have the same characteristics. And this question of adequacy is none other than the first problem we noted, that of determining successful tests for characteristics. Furthermore, unless we thought that the problem of determining success was at least <u>capable</u> of solution, then the second problem could not properly arise. If we could not in principle find a test for some characteristic, then we could never test some proposed example for that characteristic. The proposal to test some example for a property assumes that there is an adequate test for that property. Therefore to ask the second question presupposes that the first one of adequacy can be solved. Furthermore the second question of testing examples could not even arise unless it was at least in principle possible to find a test. For if we know a priori that no test could in principle be found, then the questions about testing subjects for characteristics could not arise. Therefore before we can answer any questions about building examples with some characteristics, the prior question of the possibility of finding adequate tests must be answered.

When we talk about robots and their differences from people, we are wondering whether there are some characteristics which people have that robots do not. This is clearly the second problem; the one of determining the existence of characteristics in various subjects. Similarly any discussion of the difference of the subjects which illustrate outward signs, and others which may have more characteristics, is again a question of testing some subjects to see if they have the characteristics which other given examples have. Thus to talk about robots and people, or outward signs of behaviour and consciousness, presupposes that the first question is capable of an affirmative answer. That is, it is assumed that we can find adequate tests of behaviour. In fact I don't think it would be going too far to say that the use of the conceptual device, robot, presupposes that behaviour, or examples of behaviour, can be adequately tested for. Thus to assess the opening argument which was used by Scriven, we must examine the possibility of establishing adequate tests for behaviour.

So far I have stated the problem of adequacy in terms of characteristics and tests, and now I would like to restate it in a more general form in order to show the fundamental character of this problem. When we say that some situation has a characteristic we: are, speaking more generally, using in a meaningful way, some concept to talk about the situation. The attribution of the characteristic 'smile' can be thought of as the meaningful use of the concept 'smile'. Although I by no means intend to equate use and meaning, I do take use to be conclusive evidence that the concept has a meaning. On the other hand, however, when we talk of tests we are, more accurately, talking about the results of tests which indicate the various properties of a situation. The statement of the result of some successful test is a statement saying that a given situation has been tested and found to have a certain property. So the results of a test can be considered as the statement that a given situation has a property. The question of adequacy can now be considered more generally as a problem about the relationship between the meaningful useoof a concept in some situation and the results of various tests on that situation. I shall abbreviate the statement of this problem in what follows to just the problem of the relationship between concepts and properties, but it must be remembered that I am talking about the meaningful use of a concept in some particular situation and the tests which can be done on that situation. I have used Taylor's[24] terminology in talking initially about the adequacy question in terms of tests, but this second formulation of the problem in terms of concepts is the one that Hare uses in his chapter on 'Meaning and Criterion'.[26]

I want to look at the possible relationships between the properties of given situations and the concepts used to talk about these situations. (Note: talk about does not mean, exclusively, to describe!) There are theoretically quite a number of relationships and I tend to group them under two main headings (a) logical and (b) non-logical. The logical relationships are very numerous: (i) a property (p) is necessary and sufficient for the concept (c) (ii) p is necessary for c, (iii) p is sufficient for c, (iv) some group of properties (p^n) are sufficient and necessary for c, (v) p^n is necessary for c, (vi) p^n is sufficient for c, (vii) some of a group of properties (p^{n-k}) are necessary and sufficient for c, (viii) p^{n-k} is necessary for c, (ix) p^{n-k} is sufficient for c. The general form of those relationships which are necessary and sufficient is $p^{n-k}\leftrightarrow c$ where $k<n$ and $n\geqslant 1, k\geqslant o$. Similar generalforms can be found for the necessary and for the sufficient relationships. It is therefore evident that there are, in principle, no limits to the number of logical relationships between properties and concepts. And finally those properties which satisfy or belong to one of these relationships, I shall call a criterion for that concept.

Some properties however are only normally adquate for the ascription of some concept. That is to say that when a situation contains a property, or series of properties, the concept is normally applicable. There are exceptional cases, of course, but generally we are justified in using the concept when these properties exist in some situation. The relationships between the properties and the concept is not a logical one becuase we are only normally justified in using the concept when the given situation exhibits these properties. This case may arise when the properties aee good inductive evidence for the use of the concept. Some properties may be (say) only sufficient in normal circumstances, for the application of the concept. This means that the relationships between the concept and the property is such that we are not normally justified in using a concept because of the results of a test. However because it is a sufficient relationship, we can generally conclude from the results of a test to the applicability of the concept. Furthermore, other properties may be (say) necessary, in normal situations, for the application of the concept; in which case we would be justified in concluding from the applicability of the concept to the results of some test.

So we can have properties which are, within some normal range of situations either necessary or sufficient or possibly both, for the application of some concept. However the relationships are not logical in the formal sense because we can not specify the range of normal situations, nor specify the ranges that will be normal in the future. But in normal situations the properties could be necessary or sufficient or both. These properties which are related to concepts, I call (following a modified version of Scriven)[27] indicators. There are thus as many relationships between indicators and concepts as there are with criterion, but the normal relationships are not logical. Therefore it seems that there are an unlimited number of relations between concepts and properties, and even although there are two main divisions in the types of relations, even within these types there is an unlimited number of possible relations.

The question now arises quite naturally as to what types of concepts our behavioral ones are? By behavioral concepts, I mean those concepts which we use when talking about how people behave; such as, smile, smirk, grin, and grimace, to mention only a few from the various facial expressions that people adopt. It seems to me that many of the concepts are of a normal type; that is, that there are normally justifiable indications when people are smiling, but no criterion for smiles. Granting that at least at present some of our behavioral concepts are of a normal type, it may be thought that they could all be changed to a logical type. That is changed in type; but the meanings remain the same. In this regard it is interesting to consider, as an example of type reduction a paper "Can Humans 'Feel'?" by Mr. S. Coval[28] in which he argues that our behavioral concepts may become logical types as we learn more about the human organism. He argues (roughly) that we will develop behavior concepts, like "tired" which will be identified by the cause of the condition of the human. Thus if we could determine the exact tests for the causes of a piece of behavior we would have the criterion for the use of that concept. Now this suggests two alternatives, (a) that our present normal behavioral concepts could all be made logical types by finding the tests which are criterion. But here no proof is offeredto show that this is possible in principle, and I see no reason to think that all normal type concepts could possibly be made logical types. Or (b) that if we do develop a set of logical type behavior concepts, we will have two sets which are irreducible, and I do not know what sort of standard we should use to compare them, as they are different types. Of course these remarks of mine about Mr. Coval's ideas are by no means meant as a refutation, but on the other hand I do not see why, when we are considering the relations between tests and concepts, we should tackle the question with only one relation in mind, that of logically adequate. However more importantly, it is evident from an examination of Coval's paper, just where a theory is needed in order to succeed in a reduction. The reductionist must offer either some principle of comparison between concepts which are different in type or else prove *a priori* that all concepts we presently use could be made logical in type without change in meaning. In the absence of either of these proofs, we can not conclude that all of our behavioral concepts which we now employ are reducible to a logical type. Therefore we can assume in the absence of a reductive theory, that our present behavioral concepts do not have criterion.

Where does all this leave Mr. Scriven with his mechanical robots imitating human behavior? Since a robot is a mechanical device it can be talked about entirely in terms of a logical type. Nowhere is any proof offered either by Scriven or anyone else who talks about robots that our behavioral cnocepts are all of a logical type. Until they prove that the concepts we use to talk about how humans behave can be reduced to logical type terms, then, I argue, the question of mechanical imitation cannot even arise. Every concept that applies to a a machine is of a logical type; probably even of the narrower class of logical types called necessary and sufficient. Thus if some performance or movement (or action) is to be accomplished by a mechanical device, then the performance must be describeable in logical concepts. At present we recognize, talk about, and describe human behavior using normal type concepts. But the problem of mechanical imitation can only arise when human behavior is described in logical type terms. Until it is shown that all human behavior is describeable in these type of terms, then the problem of imitation does not and cannot arise. Furthermore the opening argument about the infallibility of outward signs of consciousness does not show that consciousness is something other than behavior, it only shows that our concepts about consciousness are not of a logical type, but rather are of a normal type!

Now it becomes evident that the robot-man distinction is not meant to mark something outer vs. something inner, or separate outward visible signs from inward private feelings: but rather is meant to mark the distinction between a description of human activities in logical type and non-logical type terms. Or perhaps the robot-man distinction can be thought of as distinguishing those behavioral concepts which are logical from those which are not. Here, of course the non-logical type of concepts are those that we use to talk about consciousness. The question "What is Behavior?" has become the question "What types of concepts do we use for behavior?" and now perhaps the fly can get out of the fly bottle.

SECTION V

A TEST FOR THINKING

At the conclusion of his paper "The Imitation Game",[29] Keith Gunderson tempers some of his previous criticisms with the remarks:

> Neverthless...the general question would remain unanswered: what range of examples would satisfy the implicit criterion we use in our ordinary characterization of subjects as "those capable of thought"?
>
> A corollary: If we are to keep the question "Can machines think?" interesting, we cannot withhold a positive answer simply on the grounds that it (a machine) does not duplicate human activity in every respect. The question "Can a machine think if it can do everything a human being can do?" is not an interesting question....[30]

However I do not think that these remarks justify Mr. Gunderson in qualifying his earlier criticisms. I shall argue that the concept "capable of thought" has no logically sufficient criterion. If this is so then he need not worry about our implicit (logical) criterion for the concept.

Mr. Gunderson does not find the question about machines thinking, interesting, if we grant that machines can do everything humans do. But I should think that even if a machine could do everything, we would still have sceptical grounds for withholding our mental concepts. Machines are different from humans and different in a way that other humans do not differ from each other. Since a machine is by definition different than a human, even if a machine could do everything a human does, the question of relevance of the differences will always arise and I see no reason to rule it out *a priori* as uninteresting. When we build a machine to do everything that humans can, we use different materials to build with. Even when we build a mechanical "brain" we use different materials than those the brain is made of. And because a machine is different from a human in ways that other humans are not, the sceptic can always doubt the validity of the application of mental concepts to machines. Whether or not the sceptic is justified is another interesting question, but one that can always arise with machines despite the fact that they do everything.

Gunderson's corollary that we cannot withhold a positive answer simply on the grounds that machines do not duplicate human activity in every respect, seems to me to fail to notice this ever present sceptical ground. If we could find one activity which no machine could do and this was a mental activity, then together with the implicit scepticism, there would be good grounds for withholding a positive answer. This is the reason that some philosophers have been so impressed with Godel's theorem. Godel showed that given any particular Turing machine, he could always find a theorem which a human could prove was true but the machine could not. Thus there was at least one mental activity, i.e., proving the Godelian statement of that machine, which the machine could not do. When you couple this fact with the general differences between machines and humans (or even brains), then there are good reasons for withholding mental concepts (especially thinking) from machines.

Gunderson felt however that there was a general unanswered question; namely, what range of examples would satisfy the implicit criteria we use in our ordinary characterization of subjects as "those capable of thought?" It is evident that we use the concept "capable of thought" with some subjects in some situations and not in others. Most people understand the concept and we can use it, generally, unambiguously. That is to say, the concept has a meaning which most people comprehend. Now granting that a concept has meaning , and further that the meaning can be taught to others I should say, following Wittgenstein that there must be paradigm instances of the use of the word. There must be some situations in which the concept is used correctly and we know, generally, which situations they are. The concept has been taught to us and is taught by its use in paradigm situations. However, granting all this, it does not follow that there are criteria, either implicit or explicit, for the use of this concept. More proof must be offered than the fact that the concept is learned in order to prove that meaningful concepts have logically related criteria. Yet the attempt to find a test assumes just this point, namely, that there is some test which satisfies the criteria of the concept "capable of thought." There is however, no proof offered to show that the concept has criteria. Some people who work with computers contend that they can program a computer to do any task which any person could do. They may be quite justified in this claim. They then argue that if we show them what the subjects do when we say that they are capable of thought, then they will build a computer to do that job also. However this line of reasoning presupposes that there are a definite number of specific tasks which, when completed, the label "capable of thought" cannot in logical consistency, be withheld. But we cannot allow people to argue that because we are testing a machine, the concept must be of a specific type. Rather it can only be held that if we are ever going to be able to find a test for the application of the concept, then the concept must have criteria. However if the concept does not have criteria then we cannot find a logically sufficient test for its application.

Scriven thinks that if we refuse to apply our mental vocabulary each time they build a computer to do more human achievements, then we will be making a mistake. He says:

> "The logical trap is this: no one performatory achievement will be enough to persuade us to apply the human achievement vocabulary, but if we refuse to use this vocabulary in each case separately, on this ground, we will, perhaps wrongly, have committed ourselves to avoiding it even when all the achievements are simultaneously attained."[31]

Scriven seems to think that there are a definite number, (namely, all of them) of achievements which one does to qualify for the human-achievement vocabulary. If the number is not definite (and this does not mean the number infinite) then there is no logical trap.[32] But where is the proof that all of our human-achievement concepts are of a type that have a definite number of criteria? Scriven does not offer one, and I intend to show that none can be given. I shall argue that the concept "capable of thought" is an evaluative concept which does not have any logically sufficient set of characteristics so that no test for characteristics of people will ever be found that is logically sufficient. In order to prove this, however, I must first begin by reviewing some of the conclusions that have been reached in the analysis of evaluative language.

In the fifth chapter of The Language of Morals[33] Hare reformulates Moore's criticism of naturalism in ethics. In doing so Hare shows that any attempt to reduce our evaluative terms to the statement of a definite set of descriptive characteristics must be in principle mistaken. He states:

> Let us generalize. If P is a good picture' is held to mean the same as 'P is a picture and is C' (where C is a group of characteristics), then it will become impossible to to commend pictures for beingC: it will be possible only to say that they are C. It is important to realize that this difficulty has nothing to do with the particular example I have chosen. It is not because we have chosen the wrong defining characteristics; it is because whatever defining characteristics we choose, this objection arises, that we can no longer commend an object for posse possessing those characteristics.34
>
> (my parenthesis added)

As I said, I accept entirely Hare's proof that if we are to evaluate or commend various subjects for doing or being something, then we must have evaluative concepts which are not just equivalent to an assertion of a definite set of characteristics or properties. It is a fact that we do value and commend, and as long as we continue to , we must have value concepts. Thus in the absence of any proof <u>a priori</u> that at some time humans will stop forever to evaluate, it can be assumed that we must have evaluative concepts. Thus we must have concepts which are not equivalent to the assertion of a set of characteristics.

The question now arises as to whether or not when we say "X can think", we are making an evaluative judgement. In section <u>VIII</u> of his paper Gunderson says:

> A final point: the stance is often taken that thinking is the <u>crowning</u> capacity or achievement of the human race, and that if one denies that machines can think, one in effect assigns them to some <u>lower</u> level of achievement than that attained by human being. But one might well contend that machines can't think for they do much <u>better</u> than that.35
>
> (my italics)

If we often say that thinking is the crowning capacity, or the faculty which makes us better, then to say of someone that they think is not only to say that they have some capacity but that they are commendable(or more valuable) because they have it. If we call some capacity the crowning one we are in effect saying that whoever has this capacity is commendable because of it. And to offer a reason for commendation is simply to commend someone for the reason offered. However it cannot be denied that "X has the crowning capacity, viz. ability to think" and "X can think" are different utterances.

It is a generally accepted fact that people can think, and if someone states a fact which everyone knows, then it is generally assumed that he has some other purpose in mind. For example when I tell my wife, what she already knows, that the house is dirty, I am not just stating a fact but rather I am (say) condemning this condition of the house and thus recommending that she clean it. So if someone states that people (or some person) can think and we all generally assume this, then we take it that they have some other purpose in mind in uttering the sentence. Now when we remember that we often consider the ability to think as a reason for commending people, it is not difficult to see that on some occasions at least, the purpose in saying that someone can think is to commend them. For if it is assumed that Lois can think as it generally is and we often recommend people because they can think, then to say that Lois can think is to commend her because she can think. And I think that the sentence "X can think" has just this use of commendation on some occasions I want to emphasize that all I wish to establish is that on <u>some</u> occasions. the sentence has this use, while not denying that on other occasions the sentence has other uses. But part of the meaning of the concept, if we judge its meaning by its use, is evaluative and as such will have the characteristics which Hare noted about evaluative statements. If we accept the validity of Hare's analysis of our ordinary use of evaluative concepts, then we must conclude that the concept "capable of thought" is not equivalent to the statement of a set of characteristics about humans.

The question now arises as to whether or not there is a set of characteristics which are logically sufficient for the ascription of the concept "capable of thought?" Since Hare has shown that there is no set which is equivalent, then perhaps there is some set of properties which are sufficient for ascription. In this case we would then set up a series of tests for the properties and we would have a logically sufficient group of tests which, when a machine passed them, would force us (logically) to say that the machine was capable of thought. Gunderson seems to think that there is a set when he asks for the range of examples which would satisfy the implicit criteria of the concept. But there is no necessity that meaningful concepts have logically sufficient criteria. I argued in section IV that there was an indefinite number of relations between concepts and properties. some of which were logical and others not. Granted that these relations are conventional ones, this does not show that they must be logical. The convention could be that some set of characteristics is normally sufficient for the application of the concept, but that we allow exceptional circumstances to justify the withholding of the concept.As these circumstances can be neither specified nor forseen, it is evident, as I argued in section IV, that the relationship would not be a strict or logical one.

What is the relationship, then, between an evaluative concept and the characteristics of situations? Hare argues that if we evaluate something, then we must be prepared to evaluate something relevantly similar, the same way, or else offer a justification for not doing so. And he says that the "must" is a logical one in the sense that if one refused to similarly evaluate without offering a justification, then one would have committed a contradiction. Thus to fail to offer reasons is to violate the convention, and this, Hare argues, is to involve oneself in a logical contradiction, But this is far from showing that the conventional relation is a logical one. It shows only that if one violates or refuses to participate in this language convention (after entering it by using an evaluative concept) then one commits a logical fallacy, but the convention itself could just as easily be a normal one as a logical one. If one uses a concept which, as part of its convention, requires a justification in some cases and one subsequently refuses to acknowledge the demand for a justification, then one contradicts oneself, even if the convention is only one of a normal relation between the concept and the characteristics of the situation.

But Hare's analysis of the actual conventional relation between evaluative concepts and the various properties of situations, was that if an evaluative concept is used to (say) commend a situation then one must also commend another situation or else justify why one is withholding the commendation. That the situations are both given and numerically distinct is proof enough that there are differences between them, but the convention demands a justification for the relevance of the differences in withholding evaluation. Futhermore the same characteristic may be relevant in one situation for an evaluation and not in another situation for the same evaluation. But a convention in which some definite set of characteristics are (say) sufficient for the ascription of some concept except in exceptional circumstance, i.e., those circumstances in which justification can be found, is the type of convention I called normal in sectionIV. The convention for evaluative terms is that the terms must be reapplied or justi--fication offered for not reapplying them, which means that normally they will be used in the same situations but we allow exceptionally justified situations to be exempt. Therefore I conclude on the basis of Hare's analysis and theedistinctions I drew in section IV that evaluative terms are of a normal type. Furthermore since the concept "capable of thought" is an evaluative one, it has this non-logical relation to the characteristics of situations; so that no set of tests for the characteristics of some proposed subject could be logically sufficient for the ascription of the concept. Thus no test or set of tests, could , in principle, be found which would be logically sufficient to allow us to say "Machines can think."

It may be argued that if we eliminate the evaluative content from the concept of thinking then we shall be able to find a test. In the case where we find a computer which successfully passes this test, we will then be able to say that it thinks, remembering that this use of think is non-evaluative. There are two replies to this type of criticism. If it is thought that the application of this new concept to mechanical devices is a step forward in the problem of applying mental concepts to machines, then it is a mistake. By cutting out the troublesome part of the concept, one does not thereby make gains but rather one only saves up the trouble until later. In this repect then, to change the concept is only to by-pass the trouble until later while thinking that one is making gains. The second reply is that in considering problems connected with the concept of thinking, the only way to locate the problem is by considering our present concept and its ordinary usage. When the problem "Can machines think?" was originally proposed, it was assumed that people were wondering whether or not they could say of machines, what they say of lots of other things; namely that they can think. If the concept of thinking was not the one ordinarily used and meaning what we ordinarily mean, then what other possible meaning could it have had? How should we have been able to find any meaning for the question, if the words were not used as we use them in English? If in the solution to the problem, we change the meaning of the question, how can it be argued that the original question has been answered. Those people who change the concept have not answered the question "Can machines think?" but rather some other problem that they have invented.

Gunderson seems to have thought that his initial criticisms of the Imitation Game could be countered if the implicit criterion of the concept "capable of thought" could be found. He had argued in criticism, that the Imitation Game was only one example, and a multitude of examples were needed to apply the concept. But he thought that a set of tests could be found which, when satisfactorily completed, would be logically adequate for ascription of the concept. However I have argued against this, that there is no set of tests which are logically sufficient. Gunderson's error seems to have been that he mistook the type of concept that "capable of thought" is. He thought it was a logical type concept, whereas I have argued that it is anormal or non-logical type. By type I mean type of relationship between the concept and the properties of situations. By mistaking the type of concept, some philosophers have assumed that it had a logically sufficient test and set about finding the test (or tests). However when we understand what type of concept "aapable of thought" is, I have argued, then we can see that the search for a test is in principle futile.

SECTION VI

CONCLUSION

In conclusion, I should like to restate some of the conclusions tentatively arrived at in the preceeding sections of the paper. I have argued in section III that there is good evidence that mechanical robots are predictable in the important sense. Furthermore, in section IV, I argued that even using the idea of a robot as just a conceptual device, presupposed that certain linguistic problems had been solved which indeed have not been solved. In the proceeding section,I tried to show that we could never in principle find a test which was logically sufficient for the utterance "Machine X can think." Together I think that these conclusions add up to a rather serious critique of the general arguments advanced to show that machines can think. However I think that there are more far-reaching implications to be drawn from the work in this paper.

In order to point these implications, let us review the sources of errors that I suggested other philosophers had made. In arguing against the possibility of imitation, I showed that philosophers had made an error by failing to notice a linguistic question which the whole discussion of imitation presupposed. I then went on to illustrate the complexity of relations that could exist between a concept and the situations in which they were used. Finally, I suggested that those philosophers who were concerned with finding a test for thinking had mistaken the type of concept that "capable of thought" is. I have, in fact, been continually trying to show that the source of errors have all been of a linguistic nature.

Thus one of the more general conclusions of this paper is that far more attention must be given to language and the various linguistic problems that can arise. What is needed is a systematic method for tackling these problems of language once they have been shown to be behind many of the more traditional problems. But even though we still lack a methodology, there is a great need to focus more attention upon our language, its conventions, and concept types.

There is however another way of looking at the results of this paper. Much of my work has been in an effort to change the form of the standard problems associated with the question "Can machines think?". For example, I tried to show that the problem of trying to find a test for thinking is just the problem of deterining types of concepts. In another section, I showed that the problem of constructing a robot to imitate humans was at bottom, the problem of type reduction, i.e., the problem of changing a concept of one type to another without change of meaning. In making these changes, I have tried to show that the problems which have bothered philosophers in this area are essentially linguistic in nature; that is, all the problems can be restated as linguistic ones.

When I say that I have restated a traditional or standard problem, I do not mean that I have given a synonymous rephrasing of the problem. I mean either that the standard problem can be shown to have arisen because of a lack of careful linguistic analysis, or that the traditional problem presupposes that some linguistic theories be substantiated. Or even that it can be shown that the standard problems have as their main difficulty a confusion in types of concepts. If a traditional problem is related to a linguistic one in one of these ways, then we can change it into a problem in linguistics. As I said, I have tried to do just this restating of the problems in the area of minds and machines. The conclusion that I wish to suggest is that if the problems in this area can be restated, then that is some evidence that other problems may also be restatable in this way. I must, however, grant that this paper is not very substantial evidence to suggest the possible scope of this restating programme. It is my belief that most traditional philosohpic problems can be restated as linguistic ones. Thus another more general conclusion of this paper is to suggest the possibility of a general restatement of traditional philosophic problems.

Besides the more general conclusions, there are the specific ones in criticism of the arguments for saying that machines can think. I have argued that there is good evidence to doubt the possibility ofan imitating robot and even if this evidence were missing, the whole argument using imitating robots presupposes a linguistic difficulty which has not been answered. The analogy between men and robots, I argued, was empty, and I tried to show that no logically adequate test of thinking can be found; so that these examples of machines playing games were not conclusive but rather only persuasive evidence. As Gunderson said:

> In the end, the steam drill outlasted John Henry as a digger of railway tunnels, but that did not prove the machine had muscles; it proved that muscles were not needed for digging railway tunnels.[36]

There have been many interesting points made in the arguments for thinking machines, and these points have had the effect of making most philosophers expand their concept of what a machine is. However, aside from this merit of the arguments for the affirmative, I have argued that we are still justified in saying that machines cannot think.

FOOTNOTES

[1] A.M.Turing,"Computing Machinery and Intelligence", Mind, vol.LIX, No.236 (1950).

[2] A.R.Anderson, ed, Minds and Machines, Englewood Cliffs, New Jersey, Prentice-Hall, Inc, 1964.

[3] Ibid., pp.4-5

[4] Ibid., pp.72-97.

[5] Ibid., p.75

[6] Another way of putting the preceeding remarks is to say that the 'only if' condition in "T is in state A if and only if flip flop 36 is on", can be cashed into a finite list; such as, flip flop 1 is either on or off, flip flop 2 is either on or off, etc.

[7] A.R.Anderson, op. cit., pp.73-74.

[8] A.M.Turing, loc. cit.

[9] A.Church, Introduction to Mathematical Logic, Princeton, Princeton University Press, 1956.

[10] M.Davis, Computability and Unsolvability, New York, McGraw-Hill Book Company, Inc, 1958.

[11] A.R.Anderson, op. cit., p.74.

[12] Ibid., p.81.

[13] Ibid., p.77.

[14] Ibid., p.81.

[15] Ibid., p.82.

[16] Ibid., p.43.

[17] Ibid., p.77.

[18] Ibid., p.15.

[19] M.Davis, op. cit., Chapter 1.

[20] A.R.Anderson, op. cit., pp.43-59.

[21] Ibid., pp.43-44.

[22] Ibid., pp.31-42.

[23] Ibid., p.34

[24] C.Taylor, The Explanation of Behavior, London, Routledge

and Kegan Paul, 1964, pp.82-87.

[25] R.M.Hare, The Language of Morals, Oxford, at the Clarendon Press, 1961, pp.94-110.

[26] *Ibid*., pp.94-95.

[27] M.Scriven, "The Logic of Criterion", *The Journal of Philosophy*, vol.LVI, No.22, p.857.

[28] S.C.Coval, "Can Humans 'Feel'?", unpublished.

[29] A.R.Anderson, *op. cit*.

[30] *Ibid*., p.70.

[31] S.Hook, ed., *Dimensions of Mind*, New York, New York University Press, 1960, p.124.

[32] I suspect that Scriven also sees this. Compare his article in *The Journal of Philosophy*, *op. cit.*, p.868. I use his argument only as an example of the implied logical trap argument with testing.

[33] R.M.Hare, *op. cit.*, pp.79-93.

[34] *Ibid*., p.85.

[35] A.R.Anderson, *op.cit*., p.71.

[36] *Loc. cit*.

BIBLIOGRAPHY

Books

Anderson,A.R., ed. Minds and Machines. Englewood Cliffs, New Jersey, Prentice-Hall, Inc., 1964. (Contemporary Perspectives in Philosophy Series, eds. Joel Feinberg and W.C. Salmon, vol.1).

Chappell,V.C., ed. The Philosophy of Mind. Englewood Cliffs, New Jersey, Prentice-Hall, Inc., 1962.

Davis, M. Computability and Unsolvability. New York, McGraw-Hill Book Company, Inc., 1958.

Hare,R.M. The Language of Morals. Oxford, at the Clarendon Press, 1961.

Taylor,C. The Explanation of Behavior. London, Routledge and Kegan Paul, 1964. (International Library of Philosophy and Scientific Method, ed.A.J.Ayer).

Wittgenstein,L. Philosophical Investigations. Trans.G.E.M. Anscombe. Oxford, Basil Blackwell, 1963.

Articles

Albritton,R. "On Wittgenstein's Use of the Term "Criterion'". The Journal of Philosophy, vol.LVI, No.22, pp.845-856.

Scriven,M. "The Logic of Criterion". The Journal of Philosophy, vol.LVI, No.22, pp.857-865.

Wisdom,J.O. "A New Model for the Mind-Body Relationship". The British Journal for the Philosophy of Science, vol.2 No.8 (1951-52) pp.295-301.

Wisdom,J.O. "The Hypothesis of Cybernetics". The British Journal for the Philosophy of Science, vol.2 No.5 (1951-52), pp.1-24.