SESSION 1

PAPER 3

PROGRAMS WITH COMMON SENSE

by

DR. J. McCARTHY

BIOGRAPHICAL NOTE

John McCarthy, born at Boston, Mass. in 1927, received his B.S. degree in mathematics at the California Institute of Technology in 1948, and his Ph.D. also in mathematics at Princeton University in 1951. He is at present Assistant Professor of Communication Sciences at the Massachusetts Institute of Technology.

His present interests are in the artificial intelligence problem, automatic programming and mathematical logic. He is co-editor with Dr. C. E. Shannon of "Automatic Studies".

PROGRAMS WITH COMMON SENSE

by

JOHN McCARTHY

SUMMARY

INTERESTING work is being done in programming computers to solve problems which require a high degree of intelligence in humans. However, certain elementary verbal reasoning processes so simple that they can be carried out by any non-feeble-minded human have yet to be simulated by machine programs.

This paper will discuss programs to manipulate in a suitable formal language (most likely a part of the predicate calculus) common instrumental statements. The basic program will draw immediate conclusions from a list of premises. These conclusions will be either declarative or imperative sentences. When an imperative sentence is deduced the program takes a corresponding action. These actions may include printing sentences, moving sentences on lists, and reinitiating the basic deduction process on these lists.

Facilities will be provided for communication with humans in the system via manual intervention and display devices connected to the computer.

THE *advice taker* is a proposed program for solving problems by manipulating sentences in formal languages. The main difference between it and other programs or proposed programs for manipulating formal languages (the *Logic Theory Machine* of Newell, Simon and Shaw and the Geometry Program of Gelernter) is that in the previous programs the formal system was the subject matter but the heuristics were all embodied in the program. In this program the procedures will be described as much as possible in the language itself and, in particular, the heuristics are all so described.

The main advantages we expect the *advice taker* to have is that its behaviour will be improvable merely by making statements to it, telling it about its symbolic environment and what is wanted from it. To make these statements will require little if any knowledge of the program or the

previous knowledge of the *advice taker*. One will be able to assume that the *advice taker* will have available to it a fairly wide class of immediate logical consequences of anything it is told and its previous knowledge. This property is expected to have much in common with what makes us describe certain humans as having *common sense*. We shall therefore say that *A program has common sense if it automatically deduces for itself a sufficiently wide class of immediate consequences of anything it is told and what it already knows*.

The design of this system will be a joint project with Marvin Minsky, but Minsky is not to be held responsible for the views expressed here.

Before describing the *advice taker* in any detail, I would like to describe more fully our motivation for proceeding in this direction. Our ultimate objective is to make programs that learn from their experience as effectively as humans do. It may not be realized how far we are presently from this objective. It is not hard to make machines learn from experience to make simple changes in their behaviour of a kind which has been anticipated by the programmer. For example, Samuel has included in his checker program facilities for improving the weights the machine assigns to various factors in evaluating positions. He has also included a scheme whereby the machine remembers games it has played previously and deviates from its previous play when it finds a position which it previously lost. Suppose, however, that we wanted an improvement in behavior corresponding, say, to the discovery by the machine of the principle of the opposition in checkers. No present or presently proposed schemes are capable of discovering phenomena as abstract as this.

If one wants a machine to be able to discover an abstraction, it seems most likely that the machine must be able to represent this abstraction in some relatively simple way.

There is one known way of making a machine capable of learning arbitrary behaviour; thus to anticipate every kind of behaviour. This is to make it possible for the machine to simulate arbitrary behaviours and try them out. These behaviours may be represented either by nerve nets *(ref. 2)*, by Turing machines *(ref. 3)*, or by calculator programs *(ref. 4)*. The difficulty is twofold. First, in any of these representations the density of interesting behaviours is incredibly low. Second, and even more important, small interesting changes in behaviour expressed at a high level of abstraction do not have simple representations. It is as though the human genetic structure were represented by a set of blue-prints. Then a mutation would usually result in a wart or a failure of parts to meet, or even an ungrammatical blue-print which could not be translated into an animal at all. It is very difficult to see how the genetic representation scheme manages to be general enough to represent the great variety of animals observed and yet be such that so many interesting changes in the organism are represented by small genetic changes. The problem of how such a

representation controls the development of a fertilized egg into a mature animal is even more difficult.

In our opinion, a system which is to evolve intelligence of human order should have at least the following features:

1. All behaviours must be representable in the system. Therefore, the system should either be able to construct arbitrary automata or to program in some general purpose programming language.
2. Interesting changes in behaviour must be expressible in a simple way.
3. All aspects of behaviour except the most routine must be improvable. In particular, the improving mechanism should be improvable.
4. The machine must have or evolve concepts of partial success because on difficult problems decisive successes or failures come too infrequently.
5. The system must be able to create subroutines which can be included in procedures as units. The learning of subroutines is complicated by the fact that the effect of a subroutine is not usually good or bad in itself. Therefore, the mechanism that selects subroutines should have concepts of an interesting or powerful subroutine whose application may be good under suitable conditions.

Of the 5 points mentioned above, our work concentrates mainly on the second. We base ourselves on the idea that: *In order for a program to be capable of learning something it must first be capable of being told it.* In fact, in the early versions we shall concentrate entirely on this point and attempt to achieve a system which can be told to make a specific improvement in its behaviour with no more knowledge of its internal structure or previous knowledge than is required in order to instruct a human. Once this is achieved, we may be able to tell the *advice taker* how to learn from experience.

The main distinction between the way one programs a computer and modifies the program and the way one instructs a human or will instruct the *advice taker* is this: A machine is instructed mainly in the form of a sequence of imperative sentences; while a human is instructed mainly in declarative sentences describing the situation in which action is required together with a few imperatives that say what is wanted. We shall list the advantages of the two methods of instruction.

Advantages of Imperative Sentences

1. A procedure described in imperatives is already laid out and is carried out faster.
2. One starts with a machine in a basic state and does not assume previous knowledge on the part of the machine.

Advantages of Declarative Sentences

1. Advantage can be taken of previous knowledge.
2. Declarative sentences have logical consequences and it can be arranged that the machine will have available sufficiently simple logical consequences of what it is told and what it previously knew.
3. The meaning of declaratives is much less dependent on their order than is the case with imperatives. This makes it easier to have after-thoughts.
4. The effect of a declarative is less dependent on the previous state of the system so that less knowledge of this state is required on the part of the instructor.

The only way we know of expressing abstractions (such as the previous example of the opposition in checkers) is in language. That is why we have decided to program a system which reasons verbally.

THE CONSTRUCTION OF THE ADVICE TAKER

The *advice taker* system has the following main features:

1. There is a method of representing expressions in the computer. These expressions are defined recursively as follows: A class of entities called terms is defined and a term is an expression. A sequence of expressions is an expression. These expressions are represented in the machine by list structures *(ref. 1)*.
2. Certain of these expressions may be regarded as declarative sentences in a certain logical system which will be analogous to a universal Post canonical system. The particular system chosen will depend on programming considerations but will probably have a single rule of inference which will combine substitution for variables with modus ponens. The purpose of the combination is to avoid choking the machine with special cases of general propositions already deduced.
3. There is an *immediate deduction routine* which when given a set of premises will deduce a set of immediate conclusions. Initially, the immediate deduction routine will simply write down all one-step consequences of the premises. Later, this may be elaborated so that the routine will produce some other conclusions which may be of interest. However, this routine will not use semantic heuristics; i.e. heuristics which depend on the subject matter under discussion.

The intelligence, if any, of the advice taker will not be embodied in the immediate deduction routine. This intelligence will be embodied in the procedures which choose the lists of premises to which the immediate deduction routine is to be applied. Of course, the program should never attempt to apply the immediate deduction routine simultaneously to the list of everything it knows. This would make the deduction routine take too long.

4. Not all expressions are interpreted by the system as declarative sentences. Some are the names of entities of various kinds. Certain formulas represent *objects*. For our purposes, an entity is an object if we have something to say about it other than the things which may be deduced from the form of its name. For example, to most people, the number 3812 is not an object: they have nothing to say about it except what can be deduced from its structure. On the other hand, to most Americans the number 1776 is an object because they have filed somewhere the fact that it represents the year when the American Revolution started. In the *advice taker* each object has a *property list* in which are listed the specific things we have to say about it. Some things which can be deduced from the name of the object may be included in the property list anyhow if the deduction was actually carried out and was difficult enough so that the system does not want to carry it out again.
5. Entities other than declarative sentences which can be represented by formulas in the system are individuals, functions, and programs.
6. The program is intended to operate cyclically as follows. The immediate deduction routine is applied to a list of premises and a list of individuals. Some of the conclusions have the form of imperative sentences. These are obeyed. Included in the set of imperatives which may be obeyed is the routine which deduces and obeys.

We shall illustrate the way the *advice taker* is supposed to act by means of an example. Assume that I am seated at my desk at home and I wish to go to the airport. My car is at my home also. The solution of the problem is to walk to the car and drive the car to the airport. First, we shall give a formal statement of the premises the *advice taker* uses to draw the conclusions. Then we shall discuss the heuristics which cause the *advice taker* to assemble these premises from the totality of facts it has available. The premises come in groups, and we shall explain the interpretation of each group.

1. First, we have a predicate *"at"*. *"at(x,y)"* is a formalization of *"x is at y"*. Under this heading we have the premises

1. *at (I, desk)*
2. *at (desk, home)*
3. *at (car, home)*
4. *at (home, county)*
5. *at (airport, county)*

We shall need the fact that the relation *"at"* is transitive which might be written directly as

6. $at(x,y),\ at(y,z) \rightarrow at(x,z)$

or alternatively we might instead use the more abstract premises

6'. *transitive (at)*

and

7'. *transitive (u) → (u(x,y), u(yz,z) → u(x,z))*

from which 6. can be deduced.

2. There are two rules concerning the feasibility of walking and driving.

8. *walkable(x), at(y,x), at(z,x), at(I,y) → can(go(y,z, walking))*
9. *drivable(x), at(y,x), at(z,x), at(car,y), at(I,car) → can(go(y,z, driving))*

There are also two specific facts

10. *walkable (home)*
11. *drivable (county)*

3. Next we have a rule concerned with the properties of going.

12. *did(go(x,y,z)) → at(I,y)*

4. The problem itself is posed by the premise:

13. *want(at(I,airport))*

5. The above are all the premises concerned with the particular problem. The last group of premises are common to almost all problems of this sort. They are:

14. *(x → can(y)), (did(y) → z) → canachult(x,y,z)*

The predicate *"canachult(x,y,z)"* means that in a situation to which x applies, the action y can be performed and brings about a situation to which z applies. A sort of transitivity is described by

15. *canachult(x,y,z), canachult(z,u,v) → canachult(x,prog(y,u),v).*

Here *prog(u,v)* is the program of first carrying out u and then v. (Some kind of identification of a single action u with the one step program *prog(u)* is obviously required, but the details of how this will fit into the formalism have not yet been worked out).

The final premise is the one which causes action to be taken.

16. *x,canachult(x,prog(y,z),w), want(w) → do(y)*

The argument the *advice taker* must produce in order to solve the problem deduces the following propositions in more or less the following order:

1. *at(I,desk) → can(go(desk,car,walking))*
2. *at(I,car) → can(go(home,airport,driving))*
3. *did(go(desk,car,walking)) → at(I,car)*
4. *did(go(home,airport,driving)) → at(I,airport)*
5. *canachult(at(I,desk), go(desk,car,walking), at(I,car))*
6. *canachult(at(I,car), go(home,airport,driving), at(I,airport))*
7. *canachult(at(I,desk), program(go(desk,car,walking), go(home,airport, driving)), → at(I,airport))*
8. *do(go(desk,car,walking))*

The deduction of the last proposition initiates action.

The above proposed reasoning raises two major questions of heuristic. The first is that of how the 16 premises are collected, and the second is that of how the deduction proceeds once they are found. We cannot give complete answers to either question in the present paper; they are obviously not completely separate since some of the deductions might be made before some of the premises are collected. Let us first consider the question of where the 16 premises come from.

First of all, we assert that except for the 13th premise *(want(at(I, airport))* which sets the goal) and the 1st premise *(at(I,desk)* which we shall get from a routine which answers the question "where am I"), *all the premises can reasonably be expected to be specifically present in the memory* of a machine which has competence of human order in finding its way around. That is, none of them are so specific to the problem at hand that assuming their presence in memory constitutes an anticipation of this particular problem or of a class of problems narrower than those which any human can expect to have previously solved. We must impose this requirement if we are to be able to say that the *advice taker* exhibits *common sense.*

On the other hand, while we may reasonably assume that the premises are in memory, we still have to describe how they are assembled into a list by themselves to which the deduction routine may be applied. Tentatively, we expect the *advice taker* to proceed as follows: initially, the sentence *"want(at(I,airport))"* is on a certain list *L*, called the main list, all by itself. The program begins with an observation routine which looks at the main list and puts certain statements about the contents of this list on a list called "observations of the main list". We shall not specify at present what all the possible outputs of this observation routine are but merely say that in this case it will observe that "the only statement on *L* has the form *'want(u(x))'."* (We write this out in English because we have not yet settled on a formalism for representing statements of this kind). The "deduce and obey" routine is then applied to the combination of the "observations of the main list" list, and a list called the "standing orders list". This list is rather small and is never changed, or at least is only changed in major changes of the advice taker. The contents of the "standing orders" list has not been worked out, but what must be deduced is the extraction of certain statements from property lists. Namely, the program first looks at *"want(at(I,airport))"* and attempts to copy the statements on its property list. Let us assume that it fails in this attempt because *"want(at(I,airport))"* does not have the status of an object and hence has no property list. (One might expect that if the problem of going to the airport had arisen before, *"want(at(I, airport))"* would be an object, but this might depend on whether there were routines for generalizing previous experience that would allow something of general use to be filed under that heading). Next in order of

increasing generality the machine would see if anything were filed under *"want(at(I,x))"* which would deal with the general problem of getting somewhere. One would expect that premises 6, (or 6' and 7'), 8, 9, 12, would be so filed. There would also be the formula

$$want(at(I,x)) \rightarrow do(observe(\text{where am } I))$$

which would give us premise 1. There would also be a reference to the next higher level of abstraction in the goal statement which would cause a look at the property list of *"want(x)"*. This would give us 14, 15, and 16.

We shall not try to follow the solution further except to remark that *"want(at(I,x))"* there would be a rule that starts with the premises *"at(I,y)"* and *"want(I,x)"* and has as conclusion a search for the property list of *"go(y,x,z)"*. This would presumably fail, and then there would have to be heuristics that would initiate a search for a y such that *"at(I,y)"* and *"at(airport,y)"*. This would be done by looking on the property lists of the origin and the destination and working up. Then premise 9 would be found which has as one of its premises *at(I,car)*. A repetition of the above would find premise 8, which would complete the set of premises since the other *"at"* premises would have been found as by-products of previous searches.

We hope that the presence of the heuristic rules mentioned on the property lists where we have put them will seem plausible to the reader. It should be noticed that on the higher level of abstraction many of the statements are of the stimulus-response form. One might conjecture that division in man between conscious and unconscious thought occurs at the boundary between stimulus-response heuristics which do not have to be reasoned about but only obeyed, and the others which have to serve as premises in deductions.

We hope to formalize the heuristics in another paper before we start programming the system.

REFERENCES

1. NEWELL, A. and SIMON, H. A. Empirical Explorations of the Logic Theory Machine. A Case Study in Heuristic. *Proceedings of the Western Joint Computer Conference*, p. 218 (February, 1957).
2. MINSKY, M. L. Heuristic Aspects of the Artificial Intelligence Problem. *Lincoln Laboratory Report 34-55*. (December, 1956). (See also his paper for this conference and his Princeton Ph.D. thesis).
3. McCARTHY, J. Inversion of Functions Defined by Turing Machines. In Automata Studies, *Annals of Mathematics Study Number*.
4. FRIEDBERG, R. A Learning Machine, Part I. *IBM Journal of Research and Development*, 1958, 2, No. 1.

DISCUSSION ON THE PAPER BY DR. J. McCARTHY

PROF. Y. BAR-HILLEL: Dr. McCarthy's paper belongs in the Journal of Half-Baked Ideas, the creation of which was recently proposed by Dr. I. J. Good. Dr. McCarthy will probably be the first to admit this. Before he goes on to bake his ideas fully, it might be well to give him some advice and raise some objections. He himself mentions some possible objections, but I do not think that he treats them with the full consideration they deserve; there are others he does not mention.

For lack of time, I shall not go into the first part of his paper, although I think that it contains a lot of highly unclear philosophical, or pseudo-philosophical assumptions. I shall rather spend my time in commenting on the example he works out in his paper at some length. Before I start, let me voice my protest against the general assumption of Dr. McCarthy - slightly caricatured - that a machine, if only its programme is specified with a suficient degree of carelessness, will be able to carry out satisfactorily even rather difficult tasks.

Consider the assumption that the relation he designates by "*at*" is transitive (page 81). However, since he takes both "*at(I, desk)*" and "*at(desk, home)*" as premises, I presume - though this is never made quite clear - that "at" means something like being-a-physical-part-or-in-the-immediate-spatial-neighborhood-of. But then the relation is clearly not transitive. If A is in the immediate spatial neighborhood of B and B in the immediate spatial neighborhood of C, then A need not be in the immediate spatial neighborhood of C. Otherwise, everything would turn out to be in the immediate spatial neighborhood of everything, which is surely not Dr. McCarthy's intention. Of course, starting from false premises, one can still arrive at right conclusions. We do such things quite often, and a machine could do it. But it would probably be bad advice to allow a machine to do such things consistently.

Many of the other 23 steps in Dr. McCarthy's argument are equally or more questionable, but I don't think we should spend our time showing this in detail. My major question is the following: On page 83 McCarthy states that a machine which has a competence of human order in finding its way around will have almost all the premises of the argument stored in its memory. I am at a complete loss to understand the point of this remark. If Dr. McCarthy wants to say no more than that a machine, in order to behave like a human being, must have the knowledge of a human being, then this is

surely not a very important remark to make. But if not, what was the intention of this remark?

The decisive question how a machine, even assuming that it will have somehow countless millions of facts stored in its memory, will be able to pick out those facts which will serve as premises for its deduction is promised to receive its treatment in another paper, which is quite alright for a half-baked idea.

It sounds rather incredible that the machine could have arrived at its conclusion - which, in plain English, is "Walk from your desk to your car!" - by sound deduction. This conclusion surely could not possibly follow from the premises in any serious sense. Might it not be occasionally cheaper to call a taxi and have it take you over to the airport: Couldn't you decide to cancel your flight or to do a hundred other things? I don't think it would be wise to develop a programme language so powerful as to make a machine arrive at the conclusion Dr. McCarthy apparently intends it to make.

Let me also point out that in the example the time factor has never been mentioned, probably for the sake of simplicity. But clearly this factor is here so important that it could not possibly be disregarded without distorting the whole argument. Does not the solution depend, among thousands of other things, also upon the time of my being at my desk, the time at which I have to be at the airport, the distance from the airport, the speed of my car, etc.?

To make the argument deductively sound, its complexity will have to be increased by many orders of magnitude. So long as this is not realized, any discussions of machines able to perform the deductive - and inductive! - operations necessary for treating problems of the kind brought forward by Dr. McCarthy is totally pointless. The gap between Dr. McCarthy's general programme (with which I have little quarrel, after discounting its "philosophical" features) and its execution even in such a simple case as the one discussed seems to me so enormous that much more has to be done to persuade me that even the first step in bridging this gap has already been taken.

DR. O. G. SELFRIDGE: I have a question which I think applies to this. It seems to me in much of that work, the old absolutist Prof. Bar-Hillel has really put his finger on something; he is really worried about the deduction actually made. He seemed really to worry that the system is not consistent, and he made a remark that conclusions should not be drawn from false premises. In my experience those are the only conclusions that have ever been drawn. I have never yet heard of someone drawing correct conclusions from correct premises. I mean this seriously. This, I think, is Dr. Minsky's point this morning. What this leads to is that the notion of deductive logic being something sitting there sacred which you can borrow for particularly sacred uses and producing inviolable results is a lot of nonsense.

Deductive logic is inferred as much as anything else. Most women have never inferred it, but they get on perfectly well, marrying happy husbands, raising happy children, without ever using deductive logic at all. My feeling is that my criticism of Dr. McCarthy is the other way. He assumes deductive logic, whereas in fact that is something to be concocted.

This is another important point which I think Prof. Bar-Hillel ignores in this, the criticism of the programme should not be as to whether it is logically consistent, but only will he be able to wave it around saying "this in fact works the way I want it". Dr. McCarthy would be the first to admit that his programme is not now working, so it has to be changed. Then, can you make the changes in the programme to make it work? That has nothing to do with logic. Can he amend it in such a way that it includes the logic as well as the little details of the programme? Can he manage in such a way that it works the way he does? He said at the beginning of his talk that when he makes an arbitrary change in the programme it will not work usually, and you try to fix that so that it will. He has produced at least some evidence, to me at least, that small changes in his programme will not obviously not make the programme work and might even improve it. His next point is whether he can make small changes that in fact make it work. That is what we do not know yet.

PROF. Y. BAR-HILLEL: May I ask whether you could thrash this out with Dr. McCarthy? It was my impression that Dr. McCarthy's advice taker was meant to be able, among other things, to arrive at a certain conclusion from appropriate premises by faultless deductive reasoning. If this is still his programme, then I think your defence is totally beside the point.

DR. O. G. SELFRIDGE: I am not defending his programme, I am only defending him.

DR. J. McCARTHY: Are you using the word 'programme' in the technical sense of a bunch of cards or in the sense of a project that you get money for?

PROF. Y. BAR-HILLEL: When I uttered my doubts that a machine working under the programme outlined by Dr. McCarthy would be able to do what he expects it to do, I was using "programme" in the technical sense.

DR. O. G. SELFRIDGE: In that case your criticisms are not so much philosophical as technical.

PROF. Y. BAR-HILLEL: They are purely technical. I said that I shall not make any philosophical criticisms, for lack of time.

DR. O. G. SELFRIDGE: A technical objection does not make ideas half-baked.

PROF. Y. BAR-HILLEL: A deductive argument, where you have first to find out what are the relevant premises, is something which many humans are not always able to carry out successfully. I do not see the slightest reason to believe that at present machines should be able to perform things that humans find trouble in doing. I do not think there could possibly exist a programme which would, given any problem, divide all facts in the universe into those which are and those which are not relevant for that problem. Developing such a programme seems to me to be by 10^{10} orders of magnitude more difficult than, say, the Newell-Simon problem of developing a heuristic for deduction in the propositional calculus. This cavalier way of jumping over orders of magnitude only tends to becloud the issue and throw doubt on ways of thinking for which I have a great deal of respect. By developing a powerful programme language you may have paved the way for the first step in solving problems of the kind treated in your example, but the claim of being well on the way towards their solution is a gross exaggeration. This was the major point of my objections.

DR. L. C. PAYNE: First a quick comment on the remark of no woman having ever brought up a child by means of deductive logic, the point surely is obvious. The feedback is very close: if she drops the baby in a disastrous way, she does not get another chance or she gets a great yelp. She learns very quickly by crude techniques of how to achieve precise control. There is direct feedback! If she is trying to win a spouse and tries a move which does not get the right response, she quickly changes her tack. Computer-wise, we have yet to develop an input (sensory system) and data-processing technique that can give even a gesture of such resourcefulness! It is a real-time trial and error process utilizing every bit of every nuance, quickly adapting and re-adapting.

A computer can deal with only a very small amount of information compared with the human brain, and therefore attention has to be concentrated on the efficiency with which this limited amount of information is handled. This is where one may usefully turn to deductive logic, because it will be appreciated that if a person is to benefit from all the studies and knowledge of many people in different places and epochs then synthesis of some sort is essential. Science in general is just this: it's laws subsume with great economy the mechanisms of diverse processes. For example the application of deductive logic to Newton's three laws allows us to treat of a multitude of practical applications. Hence if a computer is to have any range of activity, it must be fed with explicit rules, so that by rapid deductions or transformations of data, it can evolve a host of ramifications from a limited amount of information.

The Countess of Lovelace remarked that "a machine can originate nothing: it can only do what we order it to perform". The essence of my contention is that we can only order to perform by means of transformations of data

having as their basis existing logical systems. Because of this it might be well to summarize, perhaps boldly before an audience like this, what I think is a summary of existing logics. The first is Formal Logic and consists of statements of the form, if all A are B and C is A, then certainly C is B - the syllogistic type of statement by which one can establish direct connections. This sort of logic is reversible; that is, if you start from a given complex of consistent propositions, then it is possible to take some selection of a derivative statements and from them as a starting complex, derive propositions of the original set. If the original set contains a contradiction then all other contradictions are implied latently. The best example I know of this is one recently cited by Sir Ronald Fisher, which is said to stem from the high table at Trinity College, Cambridge. The late Professor G. H. Hardy was asked, "Do you mean to say, Hardy, that you can prove any contradiction whatsoever if you have got one contradiction?". Hardy replied, "Yes, that is so". The questioner went on, "Well, four equals five, you prove that McTaggart is the Pope". Hardy rejoined at once saying, "If four equals five, then by subtracting three from each, one equals two. McTaggart and the Pope are two, therefore McTaggart and the Pope are one"!

The other important logic is Probability Theory. This consists of statements, that if some well defined proportion of A are B, and if C is A, then only an uncertain inference in the form of a probability statement, can be made about C being B. One has to be especially careful in statements of this kind to see that the total reference set is well defined and also the sub-set having some specified attribute. This kind of consideration, treated very carefully by Sir Ronald Fisher in his "Statistical Methods and Scientific Inference", nullifies the casual attitude which, to instance an example, can remark that, "statistics means that if you take enough inaccurate statements and put them together then a more accurate statement can be made". The well defined nature of the statistical mode of reasoning, if it is respected, means you can be as logically precise as with Formal Logic, but that the kind of statement you can be logically precise about is less certain, that is, it is a probability statement.

A more restricted logic can be based on what Sir Ronald Fisher calls "mathematical likelihood". This allows quantitative statements to be made on the fullest information available; it is discussed in the reference already given.

Beyond these systems one is very suspicious of the play with random exercises which purport to produce something out of nothing. It seems to me that computers can do nothing beyond applying the existing logics to effect transformations of data, since these are the limits within which exercises can be prescribed explicitly. These limits in fact are very wide and circumscribe most of the rational procedures used by human beings. They are not limited by pure mathematics, where one is constrained to using a

limited class functions which lend themselves to analysis. Numerical solutions can certainly explore regions which would bog down a more ponderous mathematical attack. In my opinion higher mathematics can be logically precise and very penetrating only about a very small class of entities, ones that are very abstract in content. Statistics allows one to deal with a wider class with less certainty, and so on down the scale until you reach common sense, where one may be rational about fairly concrete entities. Between common sense and high mathematics one has the whole range of human rationality, but for each refinement in logic one must pay the price of dealing with more restricted classes of entities which become progressively more abstract.

DR. J. McCARTHY (in reply): Prof. Bar-Hillel has correctly observed that my paper is based on unstated philosophical assumptions although what he means by "pseudo-philosophical" is unclear. Whenever we program a computer to learn from experience we build into the programme a sort of epistemology. It might be argued that this epistemology should be made explicit before one writes the programme, but epistemology is in a foggier state than computer programming even in the present half-baked state of the latter. I hope that once we have succeeded in making computer programs reason about the world, we will be able to reformulate epistemology as a branch of applied mathematics no more mysterious or controversial than physics.

On re-reading my paper I can't see how Prof. Bar-Hillel could see in it a proposal to specify a computer programme carelessly. Since other people have proposed this as a device for achieving "creativity", I can only conclude that he had some other paper in mind.

In his criticism of my use of the symbol "*at*", Prof. Bar-Hillel seems to have misunderstood the intent of the example. First of all, I was not trying to formalize the sentence form, A is at B as it is used in English. "*at*" merely was intended to serve as a convenient mnemonic for the relation between a place and a sub-place. Second I was not proposing a practical problem for the program to solve but rather an example intended to allow us to think about the kinds of reasoning involved and how a machine may be made to perform them.

Prof. Bar-Hillel's major point concerns my statement that the premises listed could be assumed to be in memory. The intention of this statement is to explain why I have not included formalizations of statements like, "it is possible to drive from my home to the airport" among my premises. If there were n known places in the country there would be $\frac{n\ (n-1)}{2}$ such sentences and, since we are quite sure that we do not have each of them in our memories, it would be cheating to allow the machine to start with them.

The rest of Prof. Bar-Hillel's criticisms concern ways in which the model mentioned does not reflect the real world; I have already explained that this was not my intention. He is certainly right that the complexity

of the model will have to be increased for it to deal with practical problems. What we disagree on is my contention that the conceptual difficulties arise at the present level of complexity and that solving them will allow us to increase the complexity of the model easily.

With regard to the discussion between Prof. Bar-Hillel and Oliver Selfridge - The logic is intended to be faultless although its premises cannot be guaranteed. The intended conclusion is "*do(go(desk,car,walking))*" not, of course, "*at(I,airport)*". The model oversimplifies but is not intended to oversimplify to the extent of allowing one to deduce one's way to the airport.

Dr. Payne's summary of formal logic does not seem to be based on much acquaintance with it and I think he underestimates the possibilities of applying it to making machines behave intelligently.