Introductory comments by Levesque and Brachman.

15 / John McCarthy
Programs with Common Sense

This is the paper that started it all. It is not only the first paper that can clearly be seen to be about Knowledge Representation, it is one of the very first papers in all of Artificial Intelligence. Its central theme is that of a hypothetical system (called the "Advice Taker") whose performance could improve over time as a result of receiving advice, rather than by being reprogrammed. Essentially, the proposal is to construct a program that reasons deductively from a body of knowledge until it concludes that it should do certain actions, which it then performs, and the cycle repeats. The actual technical content of the paper, however, is more narrowly focused on representational issues. The most important, perhaps, is the use of a calculus of *situations*, based on first-order logic, to make statements about causality, the ability of agents, and the effect of actions. This was a very influential proposal for dealing with change in a representation system, although it is being questioned by recent work like [Hayes, Chapter 28] and [Allen, Chapter 30]. On the other hand, perhaps the main contribution of the paper is the methodology it suggests, with formal logic playing a key role in the larger enterprise. Indeed, with slogans like

> We believe that human intelligence depends essentially on the fact that we can represent in language facts about our situation, our goals, and the effects of the various actions we can perform.

and

> We base ourselves on the idea that in order for a program to be capable of learning something, it must be capable of being told it.

the paper reads somewhat like a manifesto, anticipating much of what was to follow in the field and most of the papers included in this volume.

Appeared in *Semantic Information Processing*, 403–418, edited by M. Minsky, Cambridge, MA: The MIT Press, 1968.

PROGRAMS WITH COMMON SENSE

John McCarthy
Computer Science Department
Stanford University
Stanford, CA 94305
jmc@cs.stanford.edu
http://www-formal.stanford.edu/jmc/

1959

1 Introduction

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 behavior 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.[1]

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 behavior 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

[1] 1996: This was wishful thinking. Minsky's approach to AI was quite different.

some relatively simple way.

There is one known way of making a machine capable of learning arbitrary behavior; thus to anticipate every kind of behavior. This is to make it possible for the machine to simulate arbitrary behaviors and try them out. These behaviors may be represented either by nerve nets (Minsky 1956), by Turing machines (McCarthy 1956), or by calculator programs (Friedberg 1958). The difficulty is two-fold. First, in any of these representations the density of interesting behaviors is incredibly low. Second, and even more important, small interesting changes in behavior 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 behaviors 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 behavior must be expressible in a simple way.

3. All aspects of behavior 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 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 behavior 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 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 afterthoughts.

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.

2 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 (Newell and Simon 1957).

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 "$xisaty$". Under this heading we have the premises

$$at(I, desk) \tag{1}$$

$$at(desk, home) \tag{2}$$

$$at(car, home) \tag{3}$$

$$at(home, county) \tag{4}$$

$$at(airport, county) \tag{5}$$

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

$$at(x, y), at(y, z) \rightarrow at(x, z) \tag{6}$$

or alternatively we might instead use the more abstract premises

$$transitive(at) \tag{7}$$

$$transitive(u) \rightarrow (u(x, y), u(y, z) \rightarrow u(x, z)) \tag{8}$$

from which (6) can be deduced.

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

$$walkable(x), at(y, x), at(z, x), at(I, y) \rightarrow can(go(y, z, walking)) \tag{9}$$

$$drivable(x), at(y, x), at(z, x), at(car, y), at(I, car) \rightarrow can(go(y, z, driving)) \tag{10}$$

There are also two specific facts

$$walkable(home) \tag{11}$$

$$drivable(county) \tag{12}$$

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

$$did(go(x, y, z)) \rightarrow at(I, y) \tag{13}$$

4. The problem itself is posed by the premise:

$$want(at(I, airport)) \tag{14}$$

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:

$$(x \rightarrow can(y)), (did(y) \rightarrow z) \rightarrow canachult(x, y, z) \tag{15}$$

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

$$canachult(x, y, z), canachult(z, u, v) \rightarrow canachult(x, prog(y, u), v). \tag{16}$$

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.

$$x, canachult(x, prog(y, z), w), want(w) \rightarrow do(y) \tag{17}$$

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. $at(I, desk) \rightarrow can(go(desk, car, walking))$
2. $at(I, car) \rightarrow can(go(home, airport, driving))$
3. $did(go(desk, car, walking)) \rightarrow at(I, car)$
4. $did(go(home, airport, driving)) \rightarrow at(I, airport)$
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. $canachult(at(I, desk), prog(go(desk, car, walking),$
   $go(home, airport, driving)) \rightarrow at(I, airport))$
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 17 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 17 premises came from.

First of all, we assert that except for the 14th 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 "*whereamI*", *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 has 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 (7) and (8)), (9), (10), (13), would be so filed. There would

also be the formula

$$want(at(I,x)) \rightarrow do(observe(whereamI))$$

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 (15), (16), and (17).

We shall not try to follow the solution further except to remark that on the property list of "$want(at(I,x))$" 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 (10) would be found which has as one of its premises $at(I,car)$. A repetition of the above would find premise (9), 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.

3 References

Friedberg, R. (1958). A Learning Machine, Part I *IBM Journal of Research and Development 2*, No. 1.

McCarthy, John (1956). The Inversion of Functions Defined by Turing Machines, in *Automata Studies, Annals of Mathematical Study No. 34,* Princeton, pp. 177–181.

Minsky, M.L. (1956). Heuristic Aspects of the Artificial Intellegence Problem. *Lincoln Laboratory Report*,pp.34–55.