ASCRIBING MENTAL QUALITIES TO MACHINES

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

1979

Abstract

Ascribing mental qualities like *beliefs*, *intentions* and *wants* to a machine is sometimes correct if done conservatively and is sometimes necessary to express what is known about its state. We propose some new definitional tools for this: definitions relative to an approximate theory and second order structural definitions.

1 INTRODUCTION

To ascribe certain *beliefs*, *knowledge*, *free will*, *intentions*, *consciousness*, *abilities* or *wants* to a machine or computer program is legitimate when such an ascription expresses the same information about the machine that it expresses about a person. It is useful when the ascription helps us understand the structure of the machine, its past or future behavior, or how to repair or improve it. It is perhaps never logically required even for humans, but expressing reasonably briefly what is actually known about the state of a machine in a particular situation may require ascribing mental qualities or qualities isomorphic to them [1]. Theories of belief, knowledge and wanting can be constructed for machines in a simpler setting than for humans and later applied to humans. Ascription of mental qualities is most straightforward for machines of known structure such as thermostats and computer operating systems, but is most useful when applied to entities whose structure is very incompletely known.

These views are motivated by work in artificial intelligence[2] (abbreviated AI). They can be taken as asserting that many of the philosophical problems of mind take a concrete form when one takes seriously the idea of making machines behave intelligently. In particular, AI raises for machines two issues that have heretofore been considered only in connection with people.

First, in designing intelligent programs and looking at them from the outside we need to determine the conditions under which specific mental and volitional terms are applicable. We can exemplify these problems by asking when might it be legitimate to say about a machine, *"It knows I want a reservation to Boston, and it can give it to me, but it won't"*.

Second, when we want a generally intelligent[3] computer program, we

[1](McCarthy and Hayes 1969) defines an *epistemologically adequate* representation of information as one that can express the information actually available to a subject under given circumstances. Thus when we see a person, parts of him are occluded, and we use our memory of previous looks at him and our general knowledge of humans to finish of a "picture" of him that includes both two and three dimensional information. We must also consider *metaphysically adequate* representations that can represent complete facts ignoring the subject's ability to acquire the facts in given circumstances. Thus Laplace thought that the positions and velocities of the particles in the universe gave a metaphysically adequate representation. Metaphysically adequate representations are needed for scientific and other theories, but artificial intelligence and a full philosophical treatment of common sense experience also require epistemologically adequate representations. This paper might be summarized as contending that mental concepts are needed for an epistemologically adequate representation of facts about machines, especially future intelligent machines.

[2]Work in artificial intelligence is still far from showing how to reach human-level intellectual performance. Our approach to the AI problem involves identifying the intellectual mechanisms required for problem solving and describing them precisely. Therefore we are at the end of the philosophical spectrum that requires everything to be formalized in mathematical logic. It is sometimes said that one studies philosophy in order to advance beyond one's untutored naive world-view, but unfortunately for artificial intelligence, no-one has yet been able to give a description of even a naive world-view, complete and precise enough to allow a knowledge-seeking program to be constructed in accordance with its tenets.

[3]Present AI programs operate in limited domains, e.g. play particular games, prove

must build into it a general view of what the world is like with especial attention to facts about how the information required to solve problems is to be obtained and used. Thus we must provide it with some kind of *metaphysics* (general world-view) and *epistemology* (theory of knowledge) however naive.

As much as possible, we will ascribe mental qualities separately from each other instead of bundling them in a concept of mind. This is necessary, because present machines have rather varied little minds; the mental qualities that can legitimately be ascribed to them are few and differ from machine to machine. We will not even try to meet objections like, *"Unless it also does X, it is illegitimate to speak of its having mental qualities."*

Machines as simple as thermostats can be said to have beliefs, and having beliefs seems to be a characteristic of most machines capable of problem solving performance. However, the machines mankind has so far found it useful to construct rarely have beliefs about beliefs, although such beliefs will be needed by computer programs that reason about what knowledge they lack and where to get it. Mental qualities peculiar to human-like motivational structures[4] , such as love and hate, will not be required for intelligent behavior, but we could probably program computers to exhibit them if we wanted to, because our common sense notions about them translate readily into certain program and data structures. Still other mental qualities, e.g. humor and appreciation of beauty, seem much harder to model. While we will be quite liberal in ascribing *some* mental qualities even to rather primitive machines, we will try to be conservative in our criteria for ascribing any *particular* quality.

The successive sections of this paper will give philosophical and AI reasons for ascribing beliefs to machines, two new forms of definition that seem necessary for defining mental qualities and examples of their use, examples of systems to which mental qualities are ascribed, some first attempts at defining a variety of mental qualities, some comments on other views on mental

---

theorems in a particular logical system, or understand natural language sentences covering a particular subject matter and with other semantic restrictions. General intelligence will require general models of situations changing in time, actors with goals and strategies for achieving them, and knowledge about how information can be obtained.

[4]Our opinion is that human intellectual structure is substantially determined by the intellectual problems humans face. Thus a Martian or a machine will need similar structures to solve similar problems. Dennett (1971) expresses similar views. On the other hand, the human motivational structure seems to have many accidental features that might not be found in Martians and that we would not be inclined to program into machines. This is not the place to present arguments for this viewpoint.

qualities, notes, and references.

This paper is exploratory and its presentation is non-technical. Any axioms that are presented are illustrative and not part of an axiomatic system proposed as a serious candidate for AI or philosophical use. This is regrettable for two reasons. First, AI use of these concepts requires formal axiomatization. Second, the lack of formalism focusses attention on whether the paper correctly characterizes mental qualities rather than on the formal properties of the theories proposed. I think we can attain a situation like that in the foundations of mathematics, wherein the controversies about whether to take an intuitionist or classical point of view have been mainly replaced by technical studies of intuitionist and classical theories and the relations between them. In future work, I hope to treat these matters more formally along the lines of (McCarthy 1977) and (1979). This won't eliminate controversy about the true nature of mental qualities, but I believe that their eventual resolution requires more technical knowledge than is now available.

2 WHY ASCRIBE MENTAL QUALITIES?

Why should we want to ascribe beliefs to machines at all? This is the converse question to that of reductionism. Instead of asking how mental qualities can be reduced to physical ones, we ask how to ascribe mental qualities to physical systems.

Our general motivation for ascribing mental qualities is the same as for ascribing any other qualities—namely to express available information about the machine and its current state. To have information, we must have a space of possibilities whether explicitly described or not. The ascription must therefore serve to distinguish the present state of the machine from past or future states or from the state the machine would have in other conditions or from the state of other machines. Therefore, the issue is whether ascription of mental qualities is helpful in making these discriminations in the case of machines.

To put the issue sharply, consider a computer program for which we possess complete listings. The behavior of the program in any environment is determined from the structure of the program and can be found out by simulating the action of the program and the environment without having to deal with any concept of belief. Nevertheless, there are several reasons for ascribing belief and other mental qualities:

1. Although we may know the program, its state at a given moment is usually not directly observable, and the facts we can obtain about its current state may be more readily expressed by ascribing certain beliefs and goals than in any other way.

2. Even if we can simulate its interaction with its environment using another more comprehensive program, the simulation may be a billion times too slow. We also may not have the initial conditions of the environment or the environment's laws of motion in a suitable form, whereas it may be feasible to make a prediction of the effects of the beliefs we ascribe to the program without any computer at all.

3. Ascribing beliefs may allow deriving general statements about the program's behavior that could not be obtained from any finite number of simulations.

4. The belief and goal structures we ascribe to the program may be easier to understand than the details of program as expressed in its listing.

5. The belief and goal structure is likely to be close to the structure the designer of the program had in mind, and it may be easier to debug the program in terms of this structure than directly from the listing. In fact, it is often possible for someone to correct a fault by reasoning in general terms about the information in a program or machine, diagnosing what is wrong as a false belief, and looking at the details of the program or machine only sufficiently to determine how the false belief is represented and what mechanism caused it to arise.

6. The difference between this program and another actual or hypothetical program may best be expressed as a difference in belief structure.

All the above reasons for ascribing beliefs are epistemological; i.e. ascribing beliefs is needed to adapt to limitations on our ability to acquire knowledge, use it for prediction, and establish generalizations in terms of the elementary structure of the program. Perhaps this is the general reason for ascribing higher levels of organization to systems.

Computers give rise to numerous examples of building a higher structure on the basis of a lower and conducting subsequent analyses using the higher structure. The geometry of the electric fields in a transistor and its chemical composition give rise to its properties as an electric circuit element. Transistors are combined in small circuits and powered in standard ways to make logical elements such as ANDs, ORs, NOTs and flip-flops. Computers are designed with these logical elements to obey a desired order code; the designer usually needn't consider the properties of the transistors as circuit elements. When writing a compiler from a higher level language, one works with the order code and doesn't have to know about the ANDs and ORs; the user of the higher order language needn't know the computer's order code.

In the above cases, users of the higher level can completely ignore the lower level, because the behavior of the higher level system is completely determined by the values of the higher level variables; e.g. in order to determine the outcome of a computer program, one needn't consider the flip-flops. However, when we ascribe mental structure to humans or goals to society, we always get highly incomplete systems; the higher level behavior cannot be fully predicted from higher level observations and higher level "laws" even when the underlying lower level behavior is determinate. Moreover, at a given state of science and technology, different kinds of information can be obtained from experiment and theory building at the different levels of organization.

In order to program a computer to obtain information and co-operation from people and other machines, we will have to make it ascribe knowledge, belief, and wants to other machines and people. For example, a program that plans trips will have to ascribe knowledge to travel agents and to the airline reservation computers. It must somehow treat the information in books, perhaps by ascribing to them a passive form of knowledge. The more powerful the program in interpreting what it is told, the less it has to know about how the information it can receive is represented internally in the teller and the more its ascriptions of knowledge will look like human ascriptions of knowledge to other humans.

3 TWO METHODS OF DEFINITION

AND THEIR APPLICATION TO MENTAL QUALITIES

In our opinion, a major source of problems in defining mental and intensional concepts is the weakness of the methods of definition that have been *explicitly* used. We introduce two kinds of definition: *definition relative to an approximate theory* and *second order structural definition* and apply them to defining mental qualities.

3.1. Definitions Relative to an Approximate Theory

It is commonplace that most scientific concepts are not defined by isolated sentences of natural languages but rather as parts of theories, and the acceptance of the theory is determined by its fit to a large collection of phenomena. We propose a similar method for explicating mental and other common sense concepts, but a certain phenomenon plays a more important role than with scientific theories: the concept is meaningful only in the theory, and cannot be defined with more precision than the theory permits.

The notion of one theory approximating another needs to be formalized. In the case of physics, one can think of various kinds of numerical or probabilistic approximation. I think this kind of approximation is untypical and misleading and won't help explicate such concepts as *intentional action* as meaningful in approximate theories. Instead it may go something like this:

Consider a detailed theory T that has a state variable s. We may imagine that s changes with time. The approximating theory T' has a state variable s'. There is a predicate $atp(s, T')$ whose truth means that T' is applicable when the world is in state s. There is a relation $corr(s, s')$ which asserts that s' corresponds to the state s. We have

$$\forall s.(atp(s, T') \rightarrow \exists s'.corr(s, s')).$$

Certain functions $f1(s)$, $f2(s)$, etc. have corresponding functions $f1'(s')$, $f2'(s')$, etc. We have relations like

$$\forall ss'.(corr(s, s') \rightarrow f1(s) = f1'(s')).$$

However, the approximate theory T' may have additional functions $g1'(s')$, etc. that do not correspond to any functions of s. Even when it is possible to construct gs corresponding to the g's, their definitions will often seem arbitrary, because the common sense user of $g1'$ will only have used it within the context of T'. Concepts whose definition involves counterfactuals provide examples.

Suppose we want to ascribe *intentions* and *free will* and to distinguish a *deliberate action* from an occurrence. We want to call an output a *deliberate action* if the output would have been different if the machine's intentions had been different. This requires a criterion for the truth of the counterfactual conditional sentence *If its intentions had been different the output wouldn't have occurred*, and we require what seems to be a novel treatment of counterfactuals.

We treat the "relevant aspect of reality" as a Cartesian product so that we can talk about changing one component and leaving the others unchanged. This would be straightforward if the Cartesian product structure existed in the world; however, it usually exists only in certain approximate models of the world. Consequently no single definite state of the world as a whole corresponds to changing one component. The following paragraphs present these ideas in greater detail.

Suppose A is a theory in which some aspect of reality is characterized by the values of three quantities x, y and z. Let f be a function of three arguments, let u be a quantity satisfying $u = f(x, y, z)$, where $f(1, 1, 1) = 3$ and $f(2, 1, 1) = 5$. Consider a state of the model in which $x = 1$, $y = 1$ and $z = 1$. Within the theory A, the counterfactual conditional sentence "$u = 3$*, but if* x *were* 2*, then* u *would be* 5" is true, because the counterfactual condition means changing x to 2 and leaving the other variables unchanged.

Now let's go beyond the model and suppose that x, y and z are quantities depending on the state of the world. Even if $u = f(x, y, z)$ is taken as a law of nature, the counterfactual need not be taken as true, because someone might argue that if x were 2, then y would be 3 so that u might not be 5. If the theory A has a sufficiently preferred status we may take the meaning of the counterfactual in A to be its general meaning, but it may sometimes be better to consider the counterfactual as defined solely in the theory, i.e. as *syncategorematic.*

A common sense example may be helpful: Suppose a ski instructor says, *"He wouldn't have fallen if he had bent his knees when he made that turn"*, and another instructor replies, *"No, the reason he fell was that he didn't put his weight on his downhill ski"*. Suppose further that on reviewing a film, they agree that the first instructor was correct and the second mistaken. I contend that this agreement is based on their common acceptance of a theory of skiing, and that *within the theory*, the decision may well be rigorous even though no-one bothers to imagine an alternate world as much like the real world as possible but in which the student had put his weight on his downhill ski.

We suggest that this is often (I haven't yet looked for counter-examples) the common sense meaning of a counterfactual. The counterfactual has a definite meaning in a theory, because the theory has a Cartesian product structure, and the theory is sufficiently preferred that the meaning of the counterfactual in the world is taken as its meaning in the theory. This is especially likely to be true for concepts that have a natural definition in terms of counterfactuals, e.g. the concept of *deliberate action* with which we started this section.

In all cases that we know about, the theory is approximate and incomplete. Provided certain propositions are true, a certain quantity is approximately a given function of certain other quantities. The incompleteness lies in the fact that the theory doesn't predict states of the world but only certain functions of them. Thus a useful concept like deliberate action may seem to vanish if examined too closely, e.g. when we try to define it in terms of states of the world and not just in terms of certain functions of these states.

Remarks:

3.1.1. The known cases in which a concept is defined relative to an approximate theory involve counterfactuals. This may not always be the case.

3.1.2 It is important to study the nature of the approximations.

3.1.3 (McCarthy and Hayes 1969) treats the notion of *X can do Y* using a theory in which the world is regarded as a collection of interacting automata. That paper failed to note that sentences using *can* cannot necessarily be translated into single assertions about the world.

3.1.4 The attempt by old fashioned introspective psychology to analyze the mind into an interacting *will*, *intellect* and other components cannot be excluded on the methodological grounds used by behaviorists and postitivists to declare them meaningless and exclude them from science. These concepts might have precise definitions within a suitable approximate theory.

3.1.5 The above treatment of counterfactuals in which they are defined in terms of the Cartesian product structure of an approximate theory may be better than the *closest possible world* treatments discussed in (Lewis 1973). The truth-values are well defined within the approximate theories, and the theories can be justified by evidence involving phenomena not mentioned in isolated counterfactual assertions.

3.1.6 Definition relative to approximate theories may help separate questions, such as some of those concerning counterfactuals, into *internal* questions within the approximate theory and the *external* question of the justification of the theory as a whole. The internal questions are likely to be technical and have definite answers on which people can agree even if they have philosophical or scientific disagreements about the external questions.

3.2. Second Order Structural Definitions

Structural definitions of qualities are given in terms of the state of the system being described while behavioral definitions are given in terms of its actual or potential behavior[5].

If the structure of the machine is known, one can give an ad hoc *first order structural definition.* This is a predicate $B(s,p)$ where s represents a state of the machine and p represents a sentence in a suitable language, and $B(s,p)$ is the assertion that when the machine is in state s, it *believes* the sentence p. (The considerations of this paper are neutral in deciding whether to regard the object of belief as a sentence or to use a modal operator or to

[5]Behavioral definitions are often favored in philosophy. A system is defined to have a certain quality if it behaves in a certain way or is *disposed* to behave in a certain way. Their virtue is conservatism; they don't postulate internal states that are unobservable to present science and may remain unobservable. However, such definitions are awkward for mental qualities, because, as common sense suggests, a mental quality may not result in behavior, because another mental quality may prevent it; e.g. I may think you are thick-headed, but politeness may prevent my saying so. Particular difficulties can be overcome, but an impression of vagueness remains. The liking for behavioral definitions stems from caution, but I would interpret scientific experience as showing that boldness in postulating complex structures of unobserved entities—provided it is accompanied by a willingness to take back mistakes—is more likely to be rewarded by understanding of and control over nature than is positivistic timidity. It is particularly instructive to imagine a determined behaviorist trying to figure out an electronic computer. Trying to define each quality behaviorally would get him nowhere; only simultaneously postulating a complex structure including memory, arithmetic unit, control structure, and input-output would yield predictions that could be compared with experiment.

There is a sense in which operational definitions are not taken seriously even by their proposers. Suppose someone gives an operational definition of length (e.g. involving a certain platinum bar), and a whole school of physicists and philosophers becomes quite attached to it. A few years later, someone else criticizes the definition as lacking some desirable property, proposes a change, and the change is accepted. This is normal, but if the original definition expressed what they really meant by the length, they would refuse to change, arguing that the new concept may have its uses, but it isn't what they mean by "length". This shows that the concept of "length" as a property of objects is more stable than any operational definition.

Carnap has an interesting section in *Meaning and Necessity* entitled "The Concept of Intension for a Robot" in which he makes a similar point saying, "It is clear that the method of structural analysis, if applicable, is more powerful than the behavioristic method, because it can supply a general answer, and, under favorable circumstances, even a complete answer to the question of the intension of a given predicate."

The clincher for AI, however, is an "argument from design". In order to produce desired behavior in a computer program, we build certain mental qualities into its structure. This doesn't lead to behavioral characterizations of the qualities, because the particular qualities are only one of many ways we might use to get the desired behavior, and anyway the desired behavior is not always realized.

admit *propositions* as abstract objects that can be believed. The paper is written as though sentences are the objects of belief, but I have more recently come to favor propositions and discuss them in (McCarthy 1979).)

A general *first order* structural definition of belief would be a predicate $B(W, M, s, p)$ where W is the "world" in which the machine M whose beliefs are in question is situated. I do not see how to give such a definition of belief, and I think it is impossible. Therefore we turn to second order definitions[6].

A second order structural definition of belief is a second order predicate $\beta(W, M, B)$. $\beta(W, M, B)$ asserts that the first order predicate B is a "good" notion of belief for the machine M in the world W. Here "good" means that the beliefs that B ascribes to M agree with our ideas of what beliefs M would have, not that the beliefs themselves are true. The axiomatizations of belief in the literature are partial second order definitions.

In general, a second order definition gives criteria for criticizing an ascription of a quality to a system. We suggest that both our common sense and scientific usage of not-directly-observable qualities corresponds more losely to second order structural definition than to any kind of behavioral definition. Note that a second order definition cannot guarantee that there exist predicates B meeting the criterion β or that such a B is unique. Some qualities are best defined jointly with related qualities, e.g. beliefs and goals may require joint treatment.

Second order definitions criticize whole belief structures rather than individual beliefs. We can treat individual beliefs by saying that a system believes p in state s provided all "reasonably good" $B's$ satisfy $B(s, p)$. Thus we are distinguishing the "intersection" of the reasonably good $B's$.

(An analogy with cryptography may be helpful. We solve a cryptogram by making hypotheses about the structure of the cipher and about the translation of parts of the cipher text. Our solution is complete when we have "guessed" a cipher system that produces the cryptogram from a plausible plaintext message. Though we never prove that our solution is unique, two different solutions are almost never found except for very short cryptograms. In the analogy, the second order definition β corresponds to the general idea of encipherment, and B is the particular system used. While we will rarely be able to prove uniqueness, we don't expect to find two $B's$ both satisfying β).

[6]Putnam (1970) also proposes what amounts to second order definitions for psychological properties.

It seems to me that there should be a metatheorem of mathematical logic asserting that not all second order definitions can be reduced to first order definitions and further theorems characterizing those second order definitions that admit such reductions. Such technical results, if they can be found, may be helpful in philosophy and in the construction of formal scientific theories. I would conjecture that many of the informal philosophical arguments that certain mental concepts cannot be reduced to physics will turn out to be sketches of arguments that these concepts require second (or higher) order definitions.

Here is an approximate second order definition of belief. For each state s of the machine and each sentence p in a suitable language L, we assign truth to $B(s,p)$ if and only if the machine is considered to believe p when it is in state s. The language L is chosen for our convenience, and there is no assumption that the machine explicitly represents sentences of L in any way. Thus we can talk about the beliefs of Chinese, dogs, corporations, thermostats, and computer operating systems without assuming that they use English or our favorite first order language. L may or may not be the language we are using for making other assertions, e.g. we could, writing in English, systematically use French sentences as objects of belief. However, the best choice for artificial intelligence work may be to make L a subset of our "outer" language restricted so as to avoid the paradoxical self-references of (Montague 1963).

We now subject $B(s,p)$ to certain criteria; i.e. $\beta(B,W)$ is considered true provided the following conditions are satisfied:

3.2.1. The set $Bel(s)$ of beliefs, i.e. the set of $p's$ for which $B(s,p)$ is assigned true when M is in state s contains sufficiently "obvious" consequences of some of its members.

3.2.2. $Bel(s)$ changes in a reasonable way when the state changes in time. We like new beliefs to be logical or "plausible" consequences of old ones or to come in as *communications* in some language on the input lines or to be *observations*, i.e. beliefs about the environment the information for which comes in on the input lines. The set of beliefs should not change too rapidly as the state changes with time.

3.2.3. We prefer the set of beliefs to be as consistent as possible. (Admittedly, consistency is not a quantitative concept in mathematical logic—a system is either consistent or not, but it would seem that we will sometimes have to ascribe inconsistent sets of beliefs to machines and people. Our intuition says that we should be able to maintain areas of consistency in our beliefs and that it may be especially important to avoid inconsistencies in the machine's purely analytic beliefs).

3.2.4. Our criteria for belief systems can be strengthened if we identify some of the machine's beliefs as expressing goals, i.e. if we have beliefs of the form "It would be good if ...". Then we can ask that the machine's behavior be somewhat *rational*, i.e. *it does what it believes will achieve its goals*. The more of its behavior we can account for in this way, the better we will like the function $B(s, p)$. We also would like to regard internal state changes as changes in belief in so far as this is reasonable.

3.2.5. If the machine communicates, i.e. emits sentences in some language that can be interpreted as assertions, questions and commands, we will want the assertions to be among its beliefs unless we are ascribing to it a goal or subgoal that involves lying. We will be most satisfied with our belief ascription, if we can account for its communications as furthering the goals we are ascribing.

3.2.6. Sometimes we shall want to ascribe introspective beliefs, e.g. a belief that it does not know how to fly to Boston or even that it doesn't know what it wants in a certain situation.

3.2.7. Finally, we will prefer a more economical ascription B to a less economical one. The fewer beliefs we ascribe and the less they change with state consistent with accounting for the behavior and the internal state changes, the better we will like it. In particular, if $\forall sp.(B1(s, p) \rightarrow B2(s, p))$, but not conversely, and $B1$ accounts for all the state changes and outputs that $B2$ does, we will prefer $B1$ to $B2$. This insures that we will prefer to assign no beliefs to stones that don't change and don't behave. A belief predicate that applies to a family of machines is preferable to one that applies to a single machine.

The above criteria have been formulated somewhat vaguely. This would be bad if there were widely different ascriptions of beliefs to a particular machine that all met our criteria or if the criteria allowed ascriptions that differed widely from our intuitions. My present opinion is that more thought will make the criteria somewhat more precise at no cost in applicability, but that they *should* still remain rather vague, i.e. we shall want to ascribe belief in a *family* of cases. However, even at the present level of vagueness, there probably won't be radically different equally "good" ascriptions of belief for systems of practical interest. If there were, we would notice unresolvable ambiguities in our ascriptions of belief to our acquaintances.

While we may not want to pin down our general idea of belief to a single axiomatization, we will need to build precise axiomatizations of belief and other mental qualities into particular intelligent computer programs.

4 EXAMPLES OF SYSTEMS WITH MENTAL QUALITIES

Let us consider some examples of machines and programs to which we may ascribe belief and goal structures.

4.1. Thermostats

Ascribing beliefs to simple thermostats is unnecessary for the study of thermostats, because their operation can be well understood without it. However, their very simplicity makes it clearer what is involved in the ascription, and we maintain (partly as a provocation to those who regard attribution of beliefs to machines as mere intellectual sloppiness) that the ascription is legitimate[7].

First consider a simple thermostat that turns off the heat when the temperature is a degree above the temperature set on the thermostat, turns on the heat when the temperature is a degree below the desired temperature, and leaves the heat as is when the temperature is in the two degree range around the desired temperature. The simplest belief predicate $B(s, p)$ ascribes belief to only three sentences: "The room is too cold", "The room is too hot", and "The room is OK"—the beliefs being assigned to states of the thermostat in the obvious way. We ascribe to it the goal, "The room should be ok". When the thermostat believes the room is too cold or too hot, it sends a message saying so to the furnace. A slightly more complex belief predicate could also be used in which the thermostat has a belief about what the temperature should be and another belief about what it is. It is not clear which is better, but if we wished to consider possible errors in the thermometer, then we would ascribe beliefs about what the temperature is. We do not ascribe to it any other beliefs; it has no opinion even about whether

[7]Whether a system has beliefs and other mental qualities is not primarily a matter of complexity of the system. Although cars are more complex than thermostats, it is hard to ascribe beliefs or goals to them, and the same is perhaps true of the basic hardware of a computer, i.e. the part of the computer that executes the program without the program itself.

the heat is on or off or about the weather or about who won the battle of Waterloo. Moreover, it has no introspective beliefs; i.e. it doesn't believe that it believes the room is too hot.

Let us compare the above $B(s, p)$ with the criteria of the previous section. The belief structure is consistent (because all the beliefs are independent of one another), they arise from observation, and they result in action in accordance with the ascribed goal. There is no reasoning and only commands (which we have not included in our discussion) are communicated. Clearly assigning beliefs is of modest intellectual benefit in this case. However, if we consider the class of possible thermostats, then the ascribed belief structure has greater constancy than the mechanisms for actually measuring and representing the temperature.

The temperature control system in my house may be described as follows: Thermostats upstairs and downstairs tell the central system to turn on or shut off hot water flow to these areas. A central water-temperature thermostat tells the furnace to turn on or off thus keeping the central hot water reservoir at the right temperature. Recently it was too hot upstairs, and the question arose as to whether the upstairs thermostat mistakenly *believed* it was too cold upstairs or whether the furnace thermostat mistakenly *believed* the water was too cold. It turned out that neither mistake was made; the downstairs controller *tried* to turn off the flow of water but *couldn't*, because the valve was stuck. The plumber came once and found the trouble, and came again when a replacement valve was ordered. Since the services of plumbers are increasingly expensive, and microcomputers are increasingly cheap, one is led to design a temperature control system that would *know* a lot more about the thermal state of the house and its own state of health.

In the first place, while the present system *couldn't* turn off the flow of hot water upstairs, there is no reason to ascribe to it the *knowledge* that it couldn't, and *a fortiori* it had no ability to *communicate* this *fact* or to take it into account in controlling the system. A more advanced system would know whether the *actions* it *attempted* succeeded, and it would communicate failures and adapt to them. (We adapted to the failure by turning off the whole system until the whole house cooled off and then letting the two parts warm up together. The present system has the *physical capability* of doing this even if it hasn't the *knowledge* or the *will*.

While the thermostat believes "The room is too cold", there is no need to say that it understands the concept of "too cold". The internal structure of "The room is too cold" is a part of our language, not its.

Consider a thermostat whose wires to the furnace have been cut. Shall we still say that it knows whether the room is too cold? Since fixing the thermostat might well be aided by ascribing this knowledge, we would like to do so. Our excuse is that we are entitled to distinguish—in our language—the concept of a broken temperature control system from the concept of a certain collection of parts, i.e. to make intensional characterizations of physical objects.

4.2. Self-Reproducing Intelligent Configurations in a Cellular Automaton World

A *cellular automaton system* assigns a finite automaton to each point of the plane with integer co-ordinates. The state of each automaton at time $t+1$ depends on its state at time t and the states of its neighbors at time t. An early use of cellular automata was by von Neumann, who found a 27 state automaton whose cells could be initialized into a self-reproducing configuration that was also a universal computer. The basic automaton in von Neumann's system had a "resting" state 0, and a point in state 0 whose four neighbors were also in that state would remain in state 0. The initial configurations considered had all but a finite number of cells in state 0, and, of course, this property would persist although the number of non-zero cells might grow indefinitely with time.

The self-reproducing system used the states of a long strip of non-zero cells as a "tape" containing instructions to a "universal constructor" configuration that would construct a copy of the configuration to be reproduced but with each cell in a passive state that would persist as long as its neighbors were also in passive states. After the construction phase, the tape would be copied to make the tape for the new machine, and then the new system would be set in motion by activating one of its cells. The new system would then move away from its mother, and the process would start over. The purpose of the design was to demonstrate that arbitrarily complex configurations could be self-reproducing—the complexity being assured by also requiring that they be universal computers.

Since von Neumann's time, simpler basic cells admitting self-reproducing universal computers have been discovered. The simplest so far is the two state Life automaton of John Conway (Gosper 1976), and in rather full detail, (Poundstone 1984). The state of a cell at time $t+1$ is determined by its state at time t and the states of its eight neighbors at time t. Namely, a point whose state is 0 will change to state 1 if exactly three of its neighbors are in state 1. A point whose state is 1 will remain in state 1 if two or three of its neighbors are in state 1. In all other cases the state becomes or remains 0.

Although this was not Conway's reason for introducing them, Conway and Gosper have shown that self-reproducing universal computers could be built up as Life configurations. Poundstone (1984) gives a full description of the Life automaton inlcuding the universal computers and self-reproducing systems.

Consider a number of such self-reproducing universal computers operating in the Life plane, and suppose that they have been programmed to study the properties of their world and to communicate among themselves about it and pursue various goals co-operatively and competitively. Call these configurations Life robots. In some respects their intellectual and scientific problems will be like ours, but in one major respect they live in a simpler world than ours seems to be. Namely, the fundamental physics of their world is that of the life automaton, and there is no obstacle to each robot *knowing* this physics, and being able to simulate the evolution of a life configuration given the initial state. Moreover, if the initial state of the robot world is finite it can have been recorded in each robot in the beginning or else recorded on a strip of cells that the robots can read. (The infinite regress of having to describe the description is avoided by providing that the description is not separately described, but can be read *both* as a description of the world *and* as a description of itself.)

Since these robots know the initial state of their world and its laws of motion, they can simulate as much of its history as they want, assuming that each can grow into unoccupied space so as to have memory to store the states of the world being simulated. This simulation is necessarily slower than real time, so they can never catch up with the present—let alone predict the future. This is obvious if the simulation is carried out straightforwardly by updating a list of currently active cells in the simulated world according to the Life rule, but it also applies to any clever mathematical method that might predict millions of steps ahead so long as it is supposed to be applicable to all Life configurations. (Some Life configurations, e.g. static ones or ones containing single *gliders* or *cannon* can have their distant futures predicted with little computing.) Namely, if there were an algorithm for such prediction, a robot could be made that would predict its own future and then disobey the prediction. The detailed proof would be analogous to the proof of unsolvability of the halting problem for Turing machines.

Now we come to the point of this long disquisition. Suppose we wish to program a robot to be successful in the Life world in competition or co-operation with the others. Without any idea of how to give a mathematical proof, I will claim that our robot will need programs that ascribe purposes and beliefs to its fellow robots and predict how they will react to its own actions by assuming that *they will act in ways that they believe will achieve their goals*. Our robot might acquire these mental theories in several ways: First, we might design the universal machine so that they are present in the initial configuration of the world. Second, we might program it to acquire these ideas by induction from its experience and even transmit them to others through an "educational system". Third, it might derive the psychological laws from the fundamental physics of the world and its knowledge of the initial configuration. Finally, it might discover how robots are built from Life cells by doing experimental "biology".

Knowing the Life physics without some information about the initial configuration is insufficient to derive the *psychological* laws, because robots can be constructed in the Life world in an infinity of ways. This follows from the "folk theorem" that the Life automaton is universal in the sense that any cellular automaton can be constructed by taking sufficiently large squares of Life cells as the basic cell of the other automaton [8].

Men are in a more difficult intellectual position than Life robots. We don't know the fundamental physics of our world, and we can't even be sure that its fundamental physics is describable in finite terms. Even if we knew the physical laws, they seem to preclude precise knowledge of an initial state and precise calculation of its future both for quantum mechanical reasons and because the continuous functions needed to represent fields seem to involve an infinite amount of information.

This example suggests that much of human mental structure is not an accident of evolution or even of the physics of our world, but is required for

[8]Our own ability to derive the laws of higher levels of organization from knowledge of lower level laws is also limited by universality. While the presently accepted laws of physics allow only one chemistry, the laws of physics and chemistry allow many biologies, and, because the neuron is a universal computing element, an arbitrary mental structure is allowed by basic neurophysiology. Therefore, to determine human mental structure, one must make psychological experiments, *or* determine the actual anatomical structure of the brain and the information stored in it. One cannot determine the structure of the brain merely from the fact that the brain is capable of certain problem solving performance. In this respect, our position is similar to that of the Life robot.

successful problem solving behavior and must be designed into or evolved by any system that exhibits such behavior.

4.3. Computer Time-Sharing Systems

These complicated computer programs allocate computer time and other resources among users. They allow each user of the computer to behave as though he had a computer of his own, but also allow them to share files of data and programs and to communicate with each other. They are often used for many years with continual small changes, and the people making the changes and correcting errors are often different from the original authors of the system. A person confronted with the task of correcting a malfunction or making a change in a time-sharing system often can conveniently use a mentalistic model of the system.

Thus suppose a user complains that the system will not run his program. Perhaps the system believes that he doesn't want to run, perhaps it persistently believes that he has just run, perhaps it believes that his quota of computer resources is exhausted, or perhaps it believes that his program requires a resource that is unavailable. Testing these hypotheses can often be done with surprisingly little understanding of the internal workings of the program.

4.4. Programs Designed to Reason

Suppose we explicitly design a program to represent information by sentences in a certain language stored in the memory of the computer and decide what to do by making inferences, and doing what it concludes will advance its goals. Naturally, we would hope that our previous second order definition of belief will "approve of" a $B(s, p)$ that ascribed to the program believing the sentences explicitly built in. We would be somewhat embarassed if someone were to show that our second order definition approved as well or better of an entirely different set of beliefs.

Such a program was first proposed in (McCarthy 1959), and here is how it might work:

Information about the world is stored in a wide variety of data structures. For example, a visual scene received by a TV camera may be represented by a $512 \times 512 \times 3$ array of numbers representing the intensities of three colors at the points of the visual field. At another level, the same scene may be represented by a list of regions, and at a further level there may be a list of physical objects and their parts together with other information about these objects obtained from non-visual sources. Moreover, information about how to solve various kinds of problems may be represented by programs in some programming language.

However, all the above representations are subordinate to a collection of sentences in a suitable first order language that includes set theory. By subordinate, we mean that there are sentences that tell what the data structures represent and what the programs do. New sentences can arise by a variety of processes: inference from sentences already present, by computation from the data structures representing observations, and by interpreting certain inputs as communications in a one or more languages.

The construction of such a program is one of the major approaches to achieving high level artificial intelligence, and, like every other approach, it faces numerous obstacles. These obstacles can be divided into two classes—*epistemological* and *heuristic*. The epistemological problem is to determine what information about the world is to be represented in the sentences and other data structures, and the heuristic problem is to decide how the information can be used effectively to solve problems. Naturally, the problems interact, but the epistemological problem is more basic and also more relevant to our present concerns. We could regard it as solved if we knew how to express the information needed for intelligent behavior so that the solution to problems logically followed from the data. The heuristic problem of actually obtaining the solutions would remain.

The information to be represented can be roughly divided into general information about the world and information about particular situations. The formalism used to represent information about the world must be *epistemologically adequate*, i.e. it must be capable of representing the information that is actually available to the program from its sensory apparatus or can be deduced. Thus it couldn't handle available information about a cup of hot coffee if its only way of representing information about fluids was in terms of the positions and velocities of the molecules. Even the hydrodynamicist's Eulerian distributions of density, velocity, temperature and pressure would be useless for representing the information actually obtainable from a television camera. These considerations are further discussed in (McCarthy and Hayes 1969).

Here are some of the kinds of general information that will have to be represented:

4.4.1. Narrative. Events occur in space and time. Some events are extended in time. Partial information must be expressed about what events begin or end during, before and after others. Partial information about places and their spacial relations must be expressible. Sometimes dynamic information such as velocities are better known than the space-time facts in terms of which they are defined.

4.4.2. Partial information about causal systems. Quantities have values and later have different values. Causal laws relate these values.

4.4.3. Some changes are results of actions by the program and other actors. Information about the effects of actions can be used to determine what goals can be achieved in given circumstances.

4.4.4 Objects and substances have locations in space. It may be that temporal and causal facts are prior to spatial facts in the formalism.

4.4.5. Some objects are actors with beliefs, purposes and intentions.

Of course, the above English description is no substitute for an axiomatized formalism, not even for philosophy, but *a fortiori* when computer programs must be written. The main difficulties in designing such a formalism involve deciding how to express partial information. (McCarthy and Hayes 1969) uses a notion of *situation* wherein the situation is never known—only facts about situations are known. Unfortunately, the formalism is not suitable for expressing what might be known when events are taking place in parallel with unknown temporal relations. It also only treats the case in which the result of an action is a definite new situation and therefore isn't suitable for describing continuous processes.

5 GLOSSARY OF MENTAL QUALITIES

In this section we give short "definitions" for machines of a collection of mental qualities. We include a number of terms which give us difficulty with an indication of what the difficulties seem to be. We emphasize the place of these concepts in the design of intelligent robots.

5.1. Introspection and Self-Knowledge

We say that a machine introspects when it comes to have beliefs about its own mental state. A simple form of introspection takes place when a program determines whether it has certain information and if not asks for it. Often an operating system will compute a check sum of itself every few minutes to verify that it hasn't been changed by a software or hardware malfunction.

In principle, introspection is easier for computer programs than for people, because the entire memory in which programs and data are stored is available for inspection. In fact, a computer program can be made to predict how it would react to particular inputs provided it has enough free storage to perform the calculation. This situation smells of paradox, and there is one. Namely, if a program could predict its own actions in less time than it takes to carry out the action, it could refuse to do what it has predicted for itself. This only shows that self-simulation is necessarily a slow process, and this is not surprising.

However, present programs do little interesting introspection. This is just a matter of the undeveloped state of artificial intelligence; programmers don't yet know how to make a computer program look at itself in a useful way.

5.2. Consciousness and Self-Consciousness

Suppose we wish to distinguish the self-awareness of a machine, animal or person from its awareness of other things. We explicate awareness as belief in certain sentences, so in this case we are want to distinguish those sentences or those terms in the sentences that may be considered to be about the self. We also don't expect that self-consciousness will be a single property that something either has or hasn't but rather there will be many kinds of self-awareness with humans posessing many of the kinds we can imagine.

Here are some of the kinds of self-awareness:

5.2.1. Certain predicates of the situation (propositional fluents in the terminology of (McCarthy and Hayes 1969)) are directly observable in almost all situations while others often must be inferred. The almost always observable fluents may reasonably be identified with the senses. Likewise the values of certain fluents are almost always under the control of the being and can be called motor parameters for lack of a common language term. We have in mind the positions of the joints. Most motor parameters are both observable and controllable. I am inclined to regard the posession of a substantial set of such constantly observable or controllable fluents as the most primitive form of self-consciousness, but I have no strong arguments against someone who wished to require more.

5.2.2. The second level of self-consciousness requires a term I in the language denoting the self. I should belong to the class of persistent objects and some of the same predicates should be applicable to it as are applicable to other objects. For example, like other objects I has a location that can change in time. I is also visible and impenetrable like other objects. However, we don't want to get carried away in regarding a physical body as a necessary condition for self-consciousness. Imagine a distributed computer whose sense and motor organs could also be in a variety of places. We don't want to exclude it from self-consciousness by definition.

5.2.3. The third level comes when I is regarded as an actor among others. The conditions that permit I to do something are similar to the conditions that permit other actors to do similar things.

5.2.4. The fourth level requires the applicability of predicates such as *believes*, *wants* and *can* to I. Beliefs about past situations and the ability to hypothesize future situations are also required for this level.

5.3. Language and Thought

Here is a hypothesis arising from artificial intelligence concerning the relation between language and thought. Imagine a person or machine that represents information internally in a huge network. Each node of the network has references to other nodes through relations. (If the system has a variable collection of relations, then the relations have to be represented by nodes, and we get a symmetrical theory if we suppose that each node is connected to a set of pairs of other nodes). We can imagine this structure to have a long term part and also extremely temporary parts representing current *thoughts*. Naturally, each being has its own network depending on its own experience. A thought is then a temporary node currently being referenced by the mechanism of consciousness. Its meaning is determined by its references to other nodes which in turn refer to yet other nodes. Now consider the problem of communicating a thought to another being.

Its full communication would involve transmitting the entire network that can be reached from the given node, and this would ordinarily constitute the entire experience of the being. More than that, it would be necessary to also communicate the programs that take action on the basis of encountering certain nodes. Even if all this could be transmitted, the recipient would still have to find equivalents for the information in terms of its own network. Therefore, thoughts have to be translated into a public language before they can be communicated.

A language is also a network of associations and programs. However, certain of the nodes in this network (more accurately a *family* of networks, since no two people speak precisely the same language) are associated with words or set phrases. Sometimes the translation from thoughts to sentences is easy, because large parts of the private networks are taken from the public network, and there is an advantage in preserving the correspondence. However, the translation is always approximate (in sense that still lacks a technical definition), and some areas of experience are difficult to translate at all. Sometimes this is for intrinsic reasons, and sometimes because particular cultures don't use language in this area. (It is my impression that cultures differ in the extent to which information about facial appearance that can be used for recognition is verbally transmitted). According to this scheme, the "deep structure" of a publicly expressible thought is a node in the public network. It is translated into the deep structure of a sentence as a tree whose terminal nodes are the nodes to which words or set phrases are attached. This "deep structure" then must be translated into a string in a spoken or written language.

The need to use language to express thought also applies when we have to ascribe thoughts to other beings, since we cannot put the entire network into a single sentence.

5.4. Intentions

We are tempted to say that a machine *intends* to perform an action when it believes it will and also believes that it could do otherwise. However, we will resist this temptation and propose that a predicate *intends(actor,action,state)* be suitably axiomatized where one of the axioms says that the machine intends the action if it believes it will perform the action and could do otherwise. Armstrong (1968) wants to require an element of servo-mechanism in order that a belief that an action will be performed be regarded as an intention, i.e. there should be a commitment to do it one way or another. There may be good reasons to allow several versions of intention to co-exist in the same formalism.

5.5. Free Will

When we program a computer to make choices intelligently after determining its options, examining their consequences, and deciding which is most favorable or most moral or whatever, we must program it to take an attitude towards its freedom of choice essentially isomorphic to that which a human must take to his own. A program will have to take such an attitude towards another unless it knows the details of the other's construction and present state.

We can define whether a particular action was free or forced *relative to a theory* that ascribes beliefs and within which beings do what they believe will advance their goals. In such a theory, action is precipitated by a belief of the form *I should do X now.* We will say that the action was free if changing the belief to *I shouldn't do X now* would have resulted in the action not being performed. This requires that the theory of belief have sufficient Cartesian product structure so that changing a single belief is defined, but it doesn't require defining what the state of the world would be if a single belief were different.

It may be possible to separate the notion of a *free action* into a technical part and a controversial part. The technical part would define freedom relative to an approximate co-ordinate system giving the necessary Cartesian product structure. Relative to the co-ordinate system, the freedom of a particular action would be a technical issue, but people could argue about whether to accept the whole co-ordinate system.

This isn't the whole free will story, because moralists are also concerned with whether praise or blame may be attributed to a choice. The following considerations would seem to apply to any attempt to define the morality of actions in a way that would apply to machines:

5.5.1. There is unlikely to be a simple behavioral definition. Instead there would be a second order definition criticizing predicates that ascribe morality to actions.

5.5.2 The theory must contain at least one axiom of morality that is not just a statement of physical fact. Relative to this axiom, moral judgments of actions can be factual.

5.5.3. The theory of morality will presuppose a theory of belief in which statements of the form *"It believed the action would harm someone"* are defined. The theory must ascribe beliefs about others' welfare and perhaps about the being's own welfare.

5.5.4. It might be necessary to consider the machine as imbedded in some kind of society in order to ascribe morality to its actions.

5.5.5. No present machines admit such a belief structure, and no such structure may be required to make a machine with arbitrarily high intelligence in the sense of problem-solving ability.

5.5.6. It seems unlikely that morally judgable machines or machines to which rights might legitimately be ascribed should be made if and when it becomes possible to do so.

5.6. Understanding

It seems to me that understanding the concept of understanding is fundamental and difficult. The first difficulty lies in determining what the operand is. What is the "theory of relativity" in *"Pat understands the theory of relativity"*? What does "misunderstand" mean? It seems that understanding should involve knowing a certain collection of facts including the general laws that permit deducing the answers to questions. We probably want to separate understanding from issues of cleverness and creativity.

5.7. Creativity

This may be easier than "understanding" at least if we confine our attention to reasoning processes. Many problem solutions involve the introduction of entities not present in the statement of the problem. For example, proving that an 8 by 8 square board with two diagonally opposite squares removed cannot be covered by dominos each covering two adjacent squares involves introducing the colors of the squares and the fact that a domino covers two squares of opposite color. We want to regard this as a creative proof even though it might be quite easy for an experienced combinatorist.

6 OTHER VIEWS ABOUT MIND

The fundamental difference in point of view between this paper and most philosophy is that we are motivated by the problem of designing an artificial intelligence. Therefore, our attitude towards a concept like *belief* is determined by trying to decide what ways of acquiring and using beliefs will lead to intelligent behavior. Then we discover that much that one intelligence can find out about another can be expressed by ascribing beliefs to it.

A negative view of empiricism seems dictated from the apparent artificiality of designing an empiricist computer program to operate in the real world. Namely, we plan to provide our program with certain senses, but we have no way of being sure that the world in which we are putting the machine is constructable from the sense impressions it will have. Whether it will ever know some fact about the world is contingent, so we are not inclined to build into it the notion that what it can't know about doesn't exist.

The philosophical views most sympathetic to our approach are some expressed by Carnap in some of the discursive sections of (Carnap 1956).

Hilary Putnam (1961) argues that the classical mind-body problems are just as acute for machines as for men. Some of his arguments are more explicit than any given here, but in that paper, he doesn't try to solve the problems for machines.

D.M. Armstrong (1968) "attempts to show that there are no valid philosophical or logical reasons for rejecting the identification of mind and brain." He does this by proposing definitions of mental concepts in terms of the state of the brain. Fundamentally, I agree with him and think that such a program of definition can be carried out, but it seems to me that his methods for defining mental qualities as brain states are too weak even for defining properties of computer programs. While he goes beyond behavioral definitions as such, he relies on dispositional states.

This paper is partly an attempt to do what Ryle (1949) says can't be done and shouldn't be attempted—namely to define mental qualities in terms of states of a machine. The attempt is based on methods of which he would not approve; he implicitly requires first order definitions, and he implicitly requires that definitions be made in terms of the state of the world and not in terms of approximate theories.

His final view of the proper subject matter of epistemology is too narrow to help researchers in artificial intelligence. Namely, we need help in expressing those facts about the world that can be obtained in an ordinary situation by an ordinary person. The general facts about the world will enable our program to decide to call a travel agent to find out how to get to Boston.

Donald Davidson (1973) undertakes to show, *"There is no important sense in which psychology can be reduced to the physical sciences"*. He proceeds by arguing that the mental qualities of a hypothetical artificial man could not be defined physically even if we knew the details of its physical structure.

One sense of Davidson's statement does not require the arguments he gives. There are many universal computing elements—relays, neurons, gates and flip-flops, and physics tells us many ways of constructing them. Any information processing system that can be constructed of one kind of element can be constructed of any other. Therefore, physics tells us nothing about what information processes exist in nature or can be constructed. Computer science is no more reducible to physics than is psychology.

However, Davidson also argues that the mental states of an organism are not describable in terms of its physical structure, and I take this to assert also that they are not describable in terms of its construction from logical elements. I would take his arguments as showing that mental qualities don't have what I have called first order structural definitions. I don't think they apply to second order definitions.

D.C. Dennett (1971) expresses views very similar to mine about the reasons for ascribing mental qualities to machines. However, the present paper emphasizes criteria for ascribing particular mental qualities to particular machines rather than the general proposition that mental qualities may be ascribed. I think that the chess programs Dennett discusses have more limited mental structures than he seems to ascribe to them. Thus their *beliefs* almost always concern particular positions, and they *believe* almost no general propositions about chess, and this accounts for many of their weaknesses. Intuitively, this is well understood by researchers in computer game playing, and providing the program with a way of representing general facts about chess and even general facts about particular positions is a major unsolved problem. For example, no present program can represent the assertion *"Black has a backward pawn on his Q3 and white may be able to cramp black's position by putting pressure on it"*. Such a representation would require rules that permit such a statement to be derived in appropriate positions and would guide the examination of possible moves in accordance with it.

I would also distinguish between believing the laws of logic and merely using them (see Dennett, p. 95). The former requires a language that can express sentences about sentences and which contains some kind of reflection principle. Many present problem solving programs can use *modus ponens* but cannot reason about their own ability to use new facts in a way that corresponds to believing *modus ponens*.

7 NOTES

Philosophy and artificial intelligence. These fields overlap in the following way: In order to make a computer program behave intelligently, its designer must build into it a view of the world in general, apart from what they include about particular sciences. (The skeptic who doubts whether there is anything to say about the world apart from the particular sciences should try to write a computer program that can figure out how to get to Timbuktoo, taking into account not only the facts about travel in general but also facts about what people and documents have what information, and what information will be required at different stages of the trip and when and how it is to be obtained. He will rapidly discover that he is lacking a *science of common sense*, i.e. he will be unable to formally express and build into his program "what everybody knows". Maybe philosophy could be defined as an attempted *science of common sense*, or else the *science of common sense* should be a definite part of philosophy.)

Artificial intelligence has another component which philosophers have not studied, namely *heuristics*. Heuristics is concerned with: given the facts and a goal, how should it investigate the possibilities and decide what to do. On the other hand, artificial intelligence is not much concerned with aesthetics and ethics.

Not all approaches to philosophy lead to results relevant to the artificial intelligence problem. On the face of it, a philosophy that entailed the view that artificial intelligence was impossible would be unhelpful, but besides that, taking artificial intelligence seriously suggests some philosophical points of view. I am not sure that all I shall list are required for pursuing the AI goal—some of them may be just my prejudices—but here they are:

1. The relation between a world view and the world should be studied by methods akin to metamathematics in which systems are studied from the outside. In metamathematics we study the relation between a mathematical system and its models. Philosophy (or perhaps *metaphilosophy*) should study the relation between world structures and systems within them that seek knowledge. Just as the metamathematician can use any mathematical methods in this study and distinguishes the methods he uses form those being studied, so the philosopher should use all his scientific knowledge in studying philosophical systems from the outside.

   Thus the question *"How do I know?"* is best answered by studying *"How does it know"*, getting the best answer that the current state of science and philosophy permits, and then seeing how this answer stands up to doubts about one's own sources of knowledge.

2. We regard *metaphysics* as the study of the general structure of the world and *epistemology* as studying what knowledge of the world can be had by an intelligence with given opportunities to observe and experiment. We need to distinguish what can be determined about the structure

of humans and machines by scientific research over a period of time and experimenting with many individuals from what can be learned in a particular situation with particular opportunities to observe. From the AI point of view, the latter is as important as the former, and we suppose that philosophers would also consider it part of epistemology. The possibilities of reductionism are also different for theoretical and everyday epistemology. We could imagine that the rules of everyday epistemology could be deduced from a knowledge of physics and the structure of the being and the world, but we can't see how one could avoid using mental concepts in expressing knowledge actually obtained by the senses.

3. It is now accepted that the basic concepts of physical theories are far removed from observation. The human sense organs are many levels of organization removed from quantum mechanical states, and we have learned to accept the complication this causes in verifying physical theories. Experience in trying to make intelligent computer programs suggests that the basic concepts of the common sense world are also complex and not always directly accessible to observation. In particular, the common sense world is not a construct from sense data, but sense data play an important role. When a man or a computer program sees a dog, we will need both the relation between the observer and the dog and the relation between the observer and the brown patch in order to construct a good theory of the event.

4. In spirit this paper is materialist, but it is logically compatible with some other philosophies. Thus cellular automaton models of the physical world may be supplemented by supposing that certain complex configurations interact with additional automata called souls that also interact with each other. Such *interactionist dualism* won't meet emotional or spiritual objections to materialism, but it does provide a logical niche for any empirically argued belief in telepathy, communication with the dead, and such other psychic phenomena as don't require tampering with causality. (As does precognition, for example). A person who believed the alleged evidence for such phenomena and still wanted scientific explanations could model his beliefs with auxiliary automata.

8 References

Armstrong, D.M. (1968). *A Materialist Theory of the Mind*, Routledge and Kegan Paul, London and New York.

Boden, Margaret A. (1972). *Purposive Explanation in Psychology*, Harvard University Press.

Carnap, Rudolf (1956). *Meaning and Necessity*, University of Chicago Press.

Davidson, Donald (1973). The Material Mind. *Logic. Methodology and Philosophy of Science IV*, P. Suppes, L. Henkin, C. Moisil, and A. Joja (eds.), Amsterdam, North-Holland.

Dennett, D.C. (1971). Intentional Systems. *Journal of Philosophy*, Vol. 68, No. 4, Feb. 25.

Gosper, R.W. (1976). Private Communication. (Much information about Life has been printed in Martin Gardner's column in *Scientific American*, and there is a magazine called *Lifeline*).

Lewis, David (1973). *Counterfactuals*, Harvard University Press.

McCarthy, John (1959). Programs with Common Sense. *Mechanisation of Thought Processes, Volume I.* London: Her Majesty's Stationery Office. (Reprinted in this volume, pp. 000–000).

McCarthy, J. and Hayes, P.J. (1969). Some Philosophical Problems from the Standpoint of Artificial Intelligence. *Machine Intelligence 4*, pp. 463–502 (eds. Meltzer, B. and Michie, D.). Edinburgh: Edinburgh University Press. (Reprinted in B. L. Webber and N. J. Nilsson (eds.), *Readings in Artificial Intelligence*, Tioga, 1981, pp. 431–450; also in M. J. Ginsberg (ed.), *Readings in Nonmonotonic Reasoning*, Morgan Kaufmann, 1987, pp. 26–45; also in this volume, pp. 000–000.)

McCarthy, John (1977). *Circumscription—A Way of Jumping to Conclusions*, Stanford Artificial Intelligence Laboratory. (Published under the title: Circumscription—a form of nonmonotonic reasoning, *Artificial Intelligence*, Vol. 13, Numbers 1,2, April. Reprinted in B. L. Webber and N. J. Nilsson (eds.), *Readings in Artificial Intelligence*, Tioga, 1981, pp. 466–472; also in M. J. Ginsberg (ed.), *Readings in Nonmonotonic Reasoning*, Morgan Kaufmann, 1987, pp. 145–152; also in this volume, pp. 000–000.)

McCarthy, J. (1979). First Order Theories of Individual Concepts and Propositions, J.E.Hayes, D.Michie and L.I.Mikulich (eds.), *Machine Intelligence 9*, Ellis Horwood. (Reprinted in this volume, pp. 000–000.)

Montague, Richard (1963). Syntactical Treatments of Modality, with Corollaries on Reflexion Principles and Finite Axiomatizability, *Acta Philosophica Fennica 16*, pp. 153–167.

Moore, E.F. (1956). Gedanken Experiments with Sequential Machines. *Automata Studies*. Princeton University Press.

Moore, Robert C. (1975). *Reasoning from Incomplete Knowledge in a Procedural Deduction System*, M.S. Thesis, M.I.T.

Poundstone, William (1984). *The Recursive Universe*. William Morrow and Co., New York.

Putnam, Hilary (1961), Minds and Machines, in *Dimensions of Mind*, Sidney Hook (ed.), Collier Books, New York.

Putnam, Hilary (1970). On Properties, in *Essays in Honor of Carl G. Hempel*, D. Reidel Publishing Co., Dordrecht, Holland.

Ryle, Gilbert (1949). *The Concept of Mind*, Hutchinson and Company, London.

/@steam.stanford.edu:/u/ftp/jmc/ascribing.tex: begun 1996 Mar 24, latexed 1996 Mar 24 at 5:04 p.m.