Making Robots Conscious of their Mental States

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

1995 July 24

Abstract

In AI, consciousness of self consists in a program having certain kinds of facts about its own mental processes and state of mind.

We discuss what consciousness of its own mental structures a robot will need in order to operate in the common sense world and accomplish the tasks humans will give it. It's quite a lot.

Many features of human consciousness will be wanted, some will not, and some abilities not possessed by humans have already been found feasible and useful in limited contexts.

We give preliminary fragments of a logical language a robot can use to represent information about its own state of mind.

A robot will often have to conclude that it cannot decide a question on the basis of the information in memory and therefore must seek information externally. Gödel's idea of relative consistency is used to formalize non-knowledge.

Programs with the kind of consciousness discussed in this article do not yet exist, although programs with some components of it exist.

Thinking about consciousness with a view to designing it provides a new approach to some of the problems of consciousness studied by philosophers. One advantage is that it focusses on the aspects of consciousness important for intelligent behavior.

Contents

1 Introduction

In this article we discuss consciousness with the methodology of logical AI. McC89[1] contains a recent discussion of logical AI. The*Remarks* section has a little about how our ideas about consciousness might apply to other AI methodologies. However, it seems that systems that don't represent information by sentences will be limited in the amount of self-consciousness they can have.

[McCarthy, 1958] proposed programs with common sense that represent what they know about particular situations and the world in general *primarily* by sentences in some language of mathematical logic. They decide what to do *primarily* by logical reasoning, i.e. when a logical AI program does an important action, it is usually because it inferred a sentence saying it should. There will usually be other data structures and programs, and they may be very important computationally, but the main decisions of what do are made by logical reasoning from sentences explicitly present in the robot's memory. Some of the sentences may get into memory by processes that run independently of the robot's decisions, e.g. facts obtained by vision. Developments in logical AI include situation calculus in various forms, logical learning, nonmonotonic reasoning in various forms ([Brewka, 1991], [Lifschitz, 1994]), theories of concepts as objects [McCarthy, 1979b] and theories of contexts as objects [McCarthy, 1993]. [McCarthy, 1958] mentioned self-observation but wasn't specific. [Dennett, 1978],[McCarthy, 1979a] and [Newell, 1980] consider regarding some information not represented by sentences explicitly present in memory as nevertheless representing sentences or propositions believed by the system. Allen Newell called this the *logic level.* I believe he did not advocate general purpose programs that represent information primarily by sentences. [2] We do.

There have been many programs that decide what do by logical reasoning with logical sentences. However, I don't know of any that are *conscious* of their own ongoing mental processes, i.e. bring sentences *about* the sentences generated by these processes into memory *along with them.* We hope to establish in this article that some consciousness of their own mental processes will be required for robots to reach a level intelligence needed to do many of the tasks humans will want to give them. In our view, **consciousness of self, i.e. introspection, is essential for human level intelligence and not a mere epiphenomenon**. However, we need to distinguish which aspects of human consciousness need to be modelled, which human qualities need not and where AI systems can go beyond human consciousness.

For the purposes of this article a robot is a continuously acting computer

[1] http://www-formal.stanford.edu/jmc/ailogic/ailogic.html

[2] Newell, together with Herbert Simon and other collaborators used logic as a domain for AI. Here the emphasis was on programs for making proofs and not in the information represented in the logical sentences.

program interacting with the outside world and not normally stopping. What physical senses and effectors or communication channels it has are irrelevant to this discussion except as examples.

In logical AI, robot consciousness may be designed as follows. At any time a certain set of sentences are directly available for reasoning. We say these sentences are in the robot's *consciousness*. Some sentences come into consciousness by processes that operate all the time, i.e. by *involuntary subconscious processes*. Others come into *consciousness* as a result of *mental actions*, e.g. observations of its consciousness, that the robot *decides* to take. The latter are the results of *introspection*.

Here's an example of human introspection. Suppose I ask you whether the President of the United States is standing, sitting or lying down at the moment, and suppose you answer that you don't know. Suppose I then ask you to think harder about it, and you answer that no amount of thinking will help. [See [Kraus et al., 1991] for one formalization.] A certain amount of introspection is required to give this answer, and robots will need a corresponding ability if they are to decide correctly whether to think more about a question or to seek the information they require externally.

We discuss what forms of consciousness and introspection are required and how some of them may be formalized. It seems that the designer of robots has many choices to make about what features of human consciousness to include. Moreover, it is very likely that useful robots will include some introspective abilities not fully possessed by humans.

Two important features of consciousness and introspection are the ability to infer nonknowledge and the ability to do nonmonotonic reasoning.

Human-like emotional structures are possible but unnecessary for useful intelligent behavior. We will also argue that it is best not to include any that would cause people either to feel sorry for robots or to dislike them.

2 What Consciousness does a Robot Need?

In some respects it is easy to provide computer programs with more powerful introspective abilities than humans have. A computer program can inspect itself, and many programs do this in a rather trivial way by computing check sums in order to verify that they have been read into computer memory without modification.

It is easy to make available for inspection by the program the manuals for the programming language used, the manual for the computer itself and a copy of the compiler. A computer program can use this information to simulate what it would do if provided with given inputs. It can answer a question like: "Would I print "YES" in less than 1,000,000 steps for a certain input? A finite version of Turing's argument that the *halting problem* is unsolvable tells us that that a computer cannot in general answer questions about what it would do in n steps in less than n steps. If it could, we (or a computer program) could construct a program that would answer a question about what it would do in n steps and then do the opposite.

Unfortunately, these easy forms of introspection are not especially useful for intelligent behavior in many common sense information situations.

We humans have rather weak memories of the events in our lives, especially of intellectual events. The ability to remember its entire intellectual history is possible for a computer program and can be used by the program in modifying its beliefs on the basis of new inferences or observations. This may prove very powerful.

To do the tasks we will give them, a robot will need many forms of self-consciousness, i.e. ability to observe its own mental state. When we say that something is *observable*, we mean that a suitable *action* by the robot causes a sentence and possibly other data structures giving the result of the observation to appear in the robot's consciousness.

We will give tentative formulas for some of the results of observations. In this we take advantage of the ideas of [McCarthy, 1993] and give a context for each formula. This makes the formulas shorter. What *Here*, *Now* and *I* mean is determined in an outer context.

- Observing its physical body, recognizing the positions of its effectors, noticing the relation of its body to the environment and noticing the values of important internal variables, e.g. the state of its power supply and of its communication channels.

  $$\ldots : c(Here, Now, I) : lowbattery \wedge in(screwdriver, hand3) \tag{1}$$

  [No reason why the robot shouldn't have three hands.]

- Observing that it does or doesn't know the value of a certain term, e.g. observing whether it knows the telephone number of a certain person. Observing that it does know the number or that it can get it by some procedure is likely to be straightforward. [3]

[3] However, observing that it doesn't know the telephone number and cannot infer what it is involves getting around Gödel's theorem. Because, if there is any sentence that is not inferrable, a system with the usual logical rules must be consistent. Therefore, it might seem that Gödel's famous theorem that the consistency of a system strong enough for arithmetic cannot be shown within the system would preclude inferring non-knowledge except for systems too weak for arithmetic. Gödel's [Gödel, 1940] idea of *relative consistency* gets us out of the difficulty.

$$c(Now, I) : \neg know(Telephone\ Clinton) \tag{2}$$

$$c(Now, I) : \neg know\text{-}whether(Sitting\ Clinton) \tag{3}$$

Deciding that it doesn't know and cannot infer the value of a telephone number is what should motivate the robot to look in the phone book or ask someone.

- Keeping a journal of physical and intellectual events so it can refer to its past beliefs, observations and actions.
- Observing its goal structure and forming sentences about it. Notice that merely having a stack of subgoals doesn't achieve this unless the stack is observable and not merely obeyable.
- The robot may *intend* to perform a certain action. It may later infer that certain possibilities are irrelevant in view of its intentions. This requires the ability to observe intentions.
- Observing how it arrived at its current beliefs. Most of the important beliefs of the system will have been obtained by nonmonotonic reasoning, and therefore are usually uncertain. It will need to maintain a critical view of these beliefs, i.e. believe meta-sentences about them that will aid in revising them when new information warrants doing so. It will presumably be useful to maintain a pedigree for each belief of the system so that it can be revised if its logical ancestors are revised. *Reason maintenance systems* maintain the pedigrees but not in the form of sentences that can be used in reasoning. Neither do they have introspective subroutines that can observe the pedigrees and generate sentences about them.
- Not only pedigrees of beliefs but other auxiliary information should either be represented as sentences or be observable in such a way as to give rise to sentences. Thus a system should be able to answer the questions: "Why do I believe p?" or alternatively "Why don't I believe p?".
- Regarding its entire mental state up to the present as an object, i.e. a context. [McCarthy, 1993] discusses contexts as formal objects. The ability to *transcend* one's present context and think about it as an object is an important form of introspection, especially when we compare human and machine intelligence as Roger Penrose (1994) and other philosophical AI critics do.

- Knowing what goals it can currently achieve and what its choices are for action. We claim that the ability to understand one's own choices constitutes *free will*. The subject is discussed in detail in [McCarthy and Hayes, 1969].

- Besides specific information about its mental state, a robot will need general facts about mental processes, so it can plan its intellectual life.

The above are only some of the needed forms of self-consciousness. Research is needed to determine their properties and to find additional useful forms of self-consciousness.

2.1 Understanding and Awareness

We do not offer definitions of understanding and awareness. Instead we discuss which abilities related to these phenomena robots will require.

Consider fish swimming. Fish do not understand swimming in the following senses.

- A fish cannot, while not swimming, review its previous swimming performance so as to swim better next time.

- A fish cannot take instruction from a more experienced fish in how to swim better.

- A fish cannot contemplate designing a fish better adapted to certain swimming conditions than it is.

A human swimmer may understand more or less about swimming. [4]

We contend that intelligent robots will need understanding of how they do things in order to improve their behavior in ways that fish cannot. Aaron Sloman [Sloman, 1985] has also discussed understanding, making the point that understanding is not an all-or-nothing quality.

Consider a robot that swims. Besides having a program for swimming with which it can interact, a logic-based robot needs to use sentences about swimming in order to give instructions to the program and to improve it.

The *understanding* a logical robot needs then requires it to use appropriate sentences about the matter being understood. The understanding involves both getting the sentences from observation and inference and using them appropriately to decide what to do.

[4] One can understand aspects of a human activity better than the people who are good at doing it. Nadia Comenici's gymnastics coach was a large, portly man hard to imagine cavorting on a gymnastics bar. Nevertheless, he *understands* women's gymnastics well enough to have coached a world champion.

Awareness is similar. It is a process whereby appropriate sentences about the world and its own mental situation come into the robot's consciousness, usually without intentional actions. Both understanding and awareness may be present to varying degrees in natural and artificial systems. The swimming robot may understand some facts about swimming and not others, and it may be aware of some aspects of its current swimming state and not others.

3 Formalized Self-Knowledge

We assume a system in which a robot maintains its information about the world and itself primarily as a collection of sentences in a mathematical logical language. There will be other data structures where they are more compact or computationally easier to process, but they will be used by programs whose results become stored as sentences. The robot decides what to do by logical reasoning, by deduction using rules of inference and also by nonmonotonic reasoning.

We do not attempt a full formalization of the rules that determine the effects of mental actions and other events in this paper. The main reason is that we are revising our theory of events to handle concurrent events in a more modular way. There is something of this in the draft [McCarthy, 1995b].

Robot consciousness involves including among its sentences some about the robot itself and about subsets of the collection of sentences itself, e.g. the sentences that were in consciousness just previous to the introspection, or at some previous time, or the sentences about a particular subject. [5]

We say subsets in order to avoid self-reference as much as possible. References to the totality of the robot's beliefs can usually be replaced by references to the totality of its beliefs up to the present moment.

3.1 Mental Situation Calculus

The *situation calculus*, initiated in [McCarthy and Hayes, 1969], is often used for describing how actions and other events affect the world. It is convenient to regard a robot's state of mind as a component of the situation and describe how mental events give rise to new situations. (We could use a formalism with a separate mental situation affected only by mental events, but this doesn't seem

[5]Too much work concerned with self-knowledge has considered self-referential sentences and getting around their apparent paradoxes. This is mostly a distraction for AI, because human self-consciousness and the self-consciousness we need to build into robots almost never involves self-referential sentences or other self-referential linguistic constructions. A simple reference to oneself is not a self-referential linguistic construction, because it isn't done by a sentence that refers to itself.

to be advantageous.) We contemplate a system in which what *holds* is closed under deductive inference, but *knowledge* is not.

The relevant notations are:

- $holds(p, s)$ is the assertion that the proposition p holds in the situation s. We shall mainly be interested in propositions p of a mental nature.

- Among the propositions that can hold are *know* p and *believe* p, where p again denotes a proposition. Thus we can have

$$holds(know\ p, s). \tag{4}$$

- As we will shortly see, sentences like

$$holds(know\ not\ know\ p, s) \tag{5}$$

  are often useful. The sentence(5) asserts that the robot knows it doesn't know p.

- Besides knowledge of propositions we need a notation for knowledge of an *individual concept*, e.g. a telephone number. [McCarthy, 1979b] treats this in some detail. That paper has separate names for objects and concepts of objects and the argument of knowing is the latter. In that paper, the symbol *mike* denotes Mike himself, the function telephone takes a person into his telephone number. Thus *telephone mike* denotes Mike's telephone number. The symbol $Mike$ is the concept of Mike, and the function $Telephone$ takes a the concept of a person into the concept of his telephone number. Thus we distinguish between Mike's telephone number, denoted by *telephone mike* and the concept of his telephone number denoted by $Telephone\ Mike$. This enables us to say

$$holds(knows\ Telephone\ Mike, s) \tag{6}$$

  to assert knowledge of Mike's telephone number and

$$holds(know\ not\ knows\ Telephone\ Mike, s) \tag{7}$$

  to mean that the robot knows it doesn't know Mike's telephone number. The notation is somewhat ponderous, but it avoids the unwanted inference that the robot knows Mary's telephone number from the facts that her telephone number is the same as Mike's and that the robot knows Mike's telephone number.[6] Having the sentence (7) in consciousness might stimulate the robot to look in the phone book.

3.2 Mental Events, especially Actions

Mental events change the situation just as do physical events.

Here is a list of some mental events, mostly described informally.

- *learn p*. The robot learns the fact p. An obvious consequence is

$$holds(know\ p, result(learn\ p, s)) \tag{8}$$

  provided the effects are definite enough to justify the *result* formalism. More likely we'll want something like

$$occurs(learn\ p, s) \supset holds(F\ know\ p, s), \tag{9}$$

  where $occurs(event, s)$ is a *point fluent* asserting that *event* occurs (instantaneously) in situation s. $F\ p$ is the proposition that the proposition p will be true at some time in the future. The *temporal function* F is used in conjunction with the function *next* and the axiom

$$holds(F\ p, s) \supset holds(p, next(p, s)). \tag{10}$$

  Here $next(p, s)$ denotes the next situation following s in which p holds. (10) asserts that if $F\ p$ holds in s, then there is a next situation in which p holds. (This *next* is not the *next* of some temporal logic formalism.)

- The robot learning p has an effect on the rest of its knowledge. We are not yet ready to propose one of the many *belief revision* systems for this. Indeed we don't assume logical closure.

- What about an event *forget p*? Forgetting p is definitely not an event with a definite result. What we can say is

$$occurs(forget\ p, s) \supset holds(F\ not\ know\ p, s) \tag{11}$$

  In general, we shall want to treat forgetting as a side-effect of some more complex event. Suppose *foo* is the more complex event. We'll have

$$occurs(foo, s) \supset occurs(forget\ p, s) \tag{12}$$

[6] Some other formalisms give up the law of substitution in logic in order to avoid this difficulty. We find the price of having separate terms for concepts worth paying in order to retain all the resources of first order logic and even higher order logic when needed.

- The robot may decide to do action a. This has the property:

$$occurs(decide\text{-}to\text{-}do\ a, s) \supset holds(intend\text{-}to\text{-}do\ a, s). \tag{13}$$

  The distinction is that *decide* is an event, and we often don't need to reason about how long it takes. *intend-to-do* is a fluent that persists until something changes it. Some call these *point fluents* and *continuous fluents* respectively.

- The robot may decide to assume p, e.g. for the sake of argument. The effect of this action is not exactly to believe p, but maybe it involves *entering a context* (see (McCarthy 1993)) in which p holds.

- The robot may infer p from other sentences, either by deduction or by some nonmonotonic form of inference.

- The robot may see some object. One result of seeing an object may be knowing that it saw the object. So we might have

$$occurs(see\ o, s) \supset holds(F\ knows\ did\ see\ o, s). \tag{14}$$

  Formalizing other effects of seeing an object require a theory of seeing that is beyond the scope of this article.

It should be obvious to the reader that we are far from having a comprehensive list of the effects of mental events. However, I hope it is also apparent that the effects of a great variety of mental events on the mental part of a situation can be formalized. Moreover, it should be clear that useful robots will need to observe mental events and reason with facts about their effects.

Most work in logical AI has involve theories in which it can be shown that a sequence of actions will achieve a goal. There are recent extensions to concurrent action, continuous action and strategies of action. All this work applies to mental actions as well.

Mostly outside this work is reasoning leading to the conclusion that a goal cannot be achieved. Similar reasoning is involved in showing that actions are safe in the sense that a certain catastrophe cannot occur. Deriving both kinds of conclusion involves inductively inferring quantified propositions, e.g. "whatever I do the goal won't be achieved" or "whatever happens the catastrophe will be avoided." This is hard for today's automated reasoning techniques, but Reiter (199x) has made important progress.

3.3 Inferring Non-knowledge

Let p be a proposition. The proposition that the robot does not know p will be written *not know* p, and we are interested in those mental situations s in which we have $holds(not\ know\ p, s)$. If *not* p is consistent with the robot's knowledge, then we certainly want $holds(not\ know\ p, s)$.

How can we assert that the proposition *not* p is consistent with the robot's knowledge? Gödel's theorem tells us that we aren't going to do it by a formal proof using the robot's knowledge as axioms. [7] The most perfunctory approach is for a program to try to prove $holds(not\ p, s)$ from the robot's knowledge and fail. Logic programming with negation as failure does this for Horn theories.

However, we can often do better. If a person or a robot regards a certain collection of facts as all that are relevant, it suffices to find a model of these facts in which p is false. [8]

Consider asserting ignorance of the value of a numerical parameter. The simplest thing is to say that there are at least two values it could have, and therefore the robot doesn't know what it is. However, we often want more, e.g. to assert that the robot knows nothing of its value. Then we must assert that the parameter could have any value, i.e. for each possible value there are models of the relevant facts in which it has that value. Of course, complete ignorance of the values of two parameters requires that there be a model in which each pair of values is taken.

It is likely to be convenient in constructing these models to assume that arithmetic is consistent, i.e. that there are models of arithmetic. Then the set of natural numbers, or equivalently Lisp S-expressions, can be used to construct the desired models. The larger the robot's collection of theories postulated to have models, the easier it will be to show ignorance.

Making a program that reasons about models of its knowledge looks diffi-

[7] We assume that our axioms are strong enough to do symbolic computation which requires the same strength as arithmetic. I think we won't get much joy from weaker systems.

[8] A conviction of about what is relevant is responsible for a person's initial reaction to the well-known puzzle of the three activists and the bear. Three Greenpeace activists have just won a battle to protect the bears' prey, the bears being already protected. It was hard work, and they decide to go see the bears whose representatives they consider themselves to have been. They wander about with their cameras, each going his own way.

Meanwhile a bear wakes up from a long sleep very hungry and heads South. After three miles, she comes across one of the activists and eats him. She then goes three miles West, finds another activist and eats her. Three miles North he finds a third activist but is too full to eat. However, annoyed by the incessant blather, she kills the remaining activist and drags him two miles East to her starting point for a nap, certain that she and her cubs can have a snack when she wakes.

What color was the bear?

At first sight it seems that the color of the bear cannot be determined from the information given. While wrong in this case, jumping to such conclusions about what is relevant is more often than not the correct thing to do.

cult, although it may turn out to be necessary in the long run. The notion of *transcending* a context may be suitable for this.

For now it seems more straightforward to use second order logic. The idea is to write the axioms of the theory with predicate and function variables and to use existential statements to assert the existence of models. Here's a proposal.

Suppose the robot has some knowledge expressed as an axiomatic theory and it needs to infer that it cannot infer *that* President Clinton is sitting down. We immediately have a problem with Gödel's incompleteness theorem, because if the theory is inconsistent, then every sentence is inferrable, and therefore a proof of non-inferrability of any sentence implies consistency. We get around this by using another idea of Gödel's—*relative consistency.*[9]

In his [Gödel, 1940], Gödel proved that if Gödel-Bernays set theory is consistent, then it remains consistent when the axiom of choice and the continuum hypothesis are added to the axioms. He did this by supposing that set theory has a model, i.e. there is a domain and an $\in$ predicate satisfying GB. He then showed that a subset of this domain, the constructible sets, provided a model of set theory in which the axiom of choice and the continuum hypothesis are also true. Cohen proved that if set theory has any models it has models in which the axiom of choice and the continuum hypothesis are false. The Gödel and Cohen proofs are long and difficult, and we don't want our robot to go through all that to show that it doesn't know that President Clinton is sitting.

For example, suppose we have a first order theory with predicate symbols $\{P_1, \ldots, P_n, sits\}$ and let $A(P_1, \ldots, P_n, sits)$ be an axiom for the theory. The second order sentence

$$(\exists P'_1, \ldots, P'_n\ sits')A(P'_1, \ldots, P'_n, sits') \tag{15}$$

expresses the consistency of the theory, and the sentence

$$(\exists P'_1, \ldots, P'_n\ sits')(A(P'_1, \ldots, P'_n, sits') \wedge \neg sits'(Clinton, s)) \tag{16}$$

expresses the consistency of the theory with the added assertion that Clinton is not sitting in the situation s.

Then

$$(15) \supset (16) \tag{17}$$

is then the required assertion of relative consistency.

Sometimes we will want to assert relative consistency under fixed interpretations of some of the predicate symbols. This would be important when we have axioms involving these predicates but do not have formulas for them, e.g. of the form $(\forall x\ y)(P(x,y) \equiv \ldots)$. Suppose, for example, that there are three predicate symbols $(P_1, P_2, sits)$, and P_1 has a fixed interpretation, and the other two are

[9]Our approach is a variant of that used by [Kraus et al., 1991].

to be chosen so as to satisfy the axiom. Then the assertion of consistency with Clinton sitting takes the form

$$(\exists P_2' P_3') A(P_1, P_2', sits') \wedge sits'(Clinton, s). \tag{18}$$

The straightforward way of proving (18) is to find substitutions for the predicate variables P_2' and $sits'$ that make the matrix of (18) true. The most trivial case of this would be when the axiom $A(P_1, P_2, sits)$ does not actually involve the predicate $sits$, and we already have an interpretation $P_1, \ldots, P_n, sits$ in which it is satisfied. Then we can define

$$sits' = (\lambda x\ ss)(\neg(x = Clinton \wedge ss = s) \vee sits(x, ss))), \tag{19}$$

and (18) follows immediately. This just means that if the new predicate does not interact with what is already known, then the values for which it is true can be assigned arbitrarily.

3.4 Observing its Motivations

Whatever motivational structure we give to robots, they should be able to observe and reason about it. For many purposes a simple goal-subgoal structure is the right thing. However, there are some elaborations to consider.

1. There often will be auxiliary goals, e.g. curiosity. When a robot is not otherwise occupied, we will want it to work at extending its knowledge.

2. The obverse of an auxiliary goal is a constraint. Maybe shall want something like Asimov's science fiction laws of robotics, e.g. that a robot should not harm humans. In a sufficiently general way of looking at goals, achieving its other goals with the constraint of not harming humans is just an elaboration of the goal itself. However, since the same constraint will apply to the achievement of many goals, it is likely to be convenient to formalize them as a separate structure. A constraint can be used to reduce the space of achievable states before the details of the goals are considered.

3.5 Robots Should Not be Equipped with Human-like Emotions

Some authors, e.g. Sloman and Croucher [Sloman and Croucher, 1981], have argued that sufficiently intelligent robots would automatically have emotions somewhat like those of humans. We argue that it is possible to give robots human-like emotions, but it would require a special effort. Moreover, it would be a bad idea if we want to use them as servants. In order to make this argument, it is necessary to assume something, as little as possible, about human emotions. Here are some points.

1. Human reasoning operates primarily on the collection of ideas of which the person is immediately conscious.

2. Other ideas are in the background and come into consciousness by various processes.

3. Because reasoning is so often nonmonotonic, conclusions can be reached on the basis of the ideas in consciousness that would not be reached if certain additional ideas were also in consciousness. [10]

4. Human emotions influence human thought by influencing what ideas come into consciousness. For example, anger brings into consciousness ideas about the target of anger and also about ways of attacking this target.

5. Human emotions are strongly related to blood chemistry. Hormones and neurotransmitters belong to the same family of substances. The sight of something frightening puts certain substances in our blood streams, and these substances may reduce the thresholds of synapses where the dendrites have receptors for these substances. [11]

6. A design that uses environmental or internal stimuli to bring whole classes of ideas into consciousness is entirely appropriate for a lower animals. We inherit this mechanism from our animal ancestors.

7. According to these notions, paranoia, schizophrenia, depression and other mental illnesses would involve malfunctions of the chemical mechanisms that bring ideas into consciousness. A paranoid who believes the Mafia or the CIA is after him and acts accordingly can lose these ideas when he takes his medicine and regain them when he stops. Certainly his blood chemistry cannot encode complicated paranoid theories, but they can bring ideas about threats from wherever or however they are stored.

These facts suggest the following design considerations.

[10] These conclusions are true in the simplest or most standard or otherwise minimal models of the ideas taken in consciousness. The point about nonmonotonicity is absolutely critical to understanding these ideas about emotion. See, for example, [McCarthy, 1980] and [McCarthy, 1986]

[11] Admittedly referring to "reducing the thresholds of synapses" is speculative. However, it may be possible to test these ideas experimentally. There should be a fixed set of these substances and therefore definite classes of ideas that they bring in.

1. We don't want robots to bring ideas into consciousness in an uncontrolled way. Robots that are to react against people (say) considered harmful, should include such reactions in their goal structures and prioritize them together with other goals. Indeed we humans advise ourselves to react rationally to danger, insult and injury. "Panic" is our name for reacting directly to perceptions of danger rather than rationally.

2. Putting such a mechanism in a robot is certainly feasible. It could be done by maintaining some numerical variables, e.g. level of fear, in the system and making the mechanism that brings sentences into consciousness (short term memory) depend on these variables. However, human-like emotional structures are not an automatic byproduct of human-level intelligence.

3. It is also practically important to avoid making robots that are reasonable targets for either human sympathy or dislike. If robots are visibly sad, bored or angry, humans, starting with children, will react to them as persons. Then they would very likely come to occupy some status in human society. Human society is complicated enough already.

4 Remarks

1. Already [Turing, 1950] disposes of "the claim that a machine cannot be the subject of its own thought". Turing further remarks

   > By observing the results of its own behavior it can modify its own programs so as to achieve some purpose more effectively. These are possibilities of the near future rather than Utopian dreams.

   We want more than than Turing explicitly asked for. The machine should oberve its processes in action and not just the results.

2. We do not give a definition of *consciousness* or *self-consciousness* in this article. We only give some properties of the consciousness phenomenon that we want robots to have together with some ideas of how to program robots accordingly.

3. The preceding sections are not to be taken as a theory of human consciousness. We do not claim that the human brain uses sentences as its primary way of representing information. Allen Newell (1980) introduced the term *logic level* of analysis of a person or machine. The idea is that behavior can be understood as the person, animal or machine doing *what it believes will achieve its goals.* Ascribing beliefs and goals then accounts

for much of its behavior. Daniel Dennett [Dennett, 1978] first introduced this idea, and it is also discussed in [McCarthy, 1979a].

Of course, logical AI involves using actual sentences in the memory of the machine.

4. Daniel Dennett [Dennett, 1991] argues that human consciousness is not a single place in the brain with every conscious idea appearing there. I think he is right about the human brain, but I think a unitary consciousness will work quite well for robots. It would likely also work for humans, but evolution happens to have produced a brain with distributed consciousness.

5. Francis Crick [Crick, 1995] discusses how to find *neurological correlates* of consciousness in the human and animal brain. I agree with all the philosophy in his paper and wish success to him and others using neuroscience. However, after reading his book, I think the artificial intelligence approach has a good chance of achieving important results sooner. They won't be quite the same results, however.

6. What about *the unconscious?* Do we need it for robots? Very likely we will need some intermediate computational processes whose results are not appropriately included in the set of sentences we take as the *consciousness* of the robot. However, they should be observable when this is useful, i.e. sentences giving facts about these processes and their results should appear in consciousness as a result of mental actions aimed at observing them. There is no need for a full-fledged Freudian unconscious with purposes of its own.

7. Should a robot hope? In what sense might it hope? How close would this be to human hope? It seems that the answer is yes. If it hopes for various things, and enough of the hopes come true, then the robot can conclude that it is doing well, and its higher level strategy is ok. If its hopes are always disappointed, then it needs to change its higher level strategy.

   To use hopes in this way requires the self observation to remember what it hoped for.

   Sometimes a robot must also infer that other robots or people hope or did hope for certain things.

8. The syntactic form is simple enough. If p is a proposition, then $hope\ p$ is the proposition that the robot hopes for p to become true. In mental situation calculus we would write

$$holds(hope\ p, s) \tag{20}$$

to assert that in mental situation s, the robot hopes for p.

Human hopes have certain qualities that I can't decide whether we will want. Hope automatically brings into consciousness thoughts related to what a situation realizing the hope would be like. We could design our programs to do the same, but this is more automatic in the human case than might be optimal. Wishful thinking is a well-known human malfunction.

9. A robot should be able to wish that it had acted differently from the way it has done. A mental example is that the robot may have taken too long to solve a problem and might wish that it had thought of the solution immediately. This will cause it to think about how it might solve such problems in the future with less computation.

10. A human can wish that his motivations and goals were different from what he observes them to be. It would seem that a program with such a wish could just change its goals.

11. [Penrose, 1994] emphasizes that a human using a logical system is prepared to accept the proposition that the system is consistent even though it can't be inferred within the system. The human is prepared to iterate this self-confidence indefinitely. Our systems should do the same, perhaps using formalized transcendence. Programs with human capability in this respect will have to be able to regard logical systems as values of variables and infer general statements about them. We will elaborate elsewhere [McCarthy, 1995a] our disagreement with Penrose about whether the human is necessarily superior to a computer program in these respects. For now we remark only that it would be interesting if he and others of similar opinion would say where they believe the efforts outlined in this article will get stuck.

12. Penrose also argues (p. 37 et seq.) that humans have *understanding* and *awareness* and machines cannot have them. He defines them in his own way, but our usage is close enough to his so that I think we are discussing how to make programs do what he thinks they cannot do. I don't agree with those defenders of AI who claim that some computer programs already possess understanding and awareness to the same extent as people. We in AI have a lot more work to do first.

13. Programs that represent information by sentences but generate new sentences by processes that don't correspond to logical reasoning present similar problems to logical AI for introspection. Approaches to AI that don't use sentences at all need some other way of representing the results of introspection if they are to use it at all.

14. Psychologists and philosophers from Aristotle have appealed to association as the main tool of thought. It is clearly inadequate to draw conclusions. We can make sense of their ideas by regarding association as the main tool for bringing facts into consciousness, but requiring reasoning to reach conclusions.

15. Some conclusions are reached by deduction, some by nonmonotonic reasoning and some by looking for models—alternatively by reasoning in second order logic.

16. Case based reasoning. Cases are *relatively rich* objects—or maybe we should say *locally rich*.

5 Acknowledgements

This work was partly supported by ARPA (ONR) grant N00014-94-1-0775 and partly done while the author was Meyerhoff Visiting Professor at the Weizmann Institute of Science, Rehovot, Israel.

Thanks to Yoav Shoham and Aaron Sloman for email comments and to Saša Buvač, Tom Costello and Donald Michie for face-to-face comments.

6 References

This document is available via the URL:http: http://www-formal.stanford.edu/jmc/.

References

[Brewka, 1991] Brewka, G. (1991). *Nonmonotonic Reasoning: Logical Foundations of Common Sense.* Cambridge University Press.

[Crick, 1995] Crick, F. (1995). *The Astonishing Hypothesis: The Scientific Search for Soul.* Scribners.

[Dennett, 1978] Dennett, D. (1978). *Brainstorms: Philosophical Essays on Mind and Psychology.* Bradford Books/MIT Press, Cambridge.

[Dennett, 1991] Dennett, D. (1991). *Consciousness Explained.* Little, Brown and Co., Boston.

[Gödel, 1940] Gödel, K. (1940). *The Consistency of The Axiom of Choice and of the Generalized Continuum-Hypothesis with the Axioms of Set Theory.* Princeton University Press.

[Kraus et al., 1991] Kraus, S., Perlis, D., and Horty, J. (1991). Reasoning about ignorance: A note on the bush-gorbachev problem. *Fundamenta Informatica*, XV:325–332.

[Lifschitz, 1994] Lifschitz, V. (1994). Circumscription. In *Handbook of Logic in Artificial Intelligence and Logic Programming, Volume 3: Nonmonotonic Reasoning and Uncertain Reasoning", publisher = "Oxford University Press.*

[McCarthy, 1958] McCarthy, J. (1958). Programs with common sense. In *Mechanisation of Thought Processes, Proceedings of the Symposium of the National Physics Laboratory*, pages 77–84, London, U.K. Her Majesty's Stationary Office. Reprinted in McC90.

[McCarthy, 1979a] McCarthy, J. (1979a). Ascribing mental qualities to machines. In Ringle, M., editor, *Philosophical Perspectives in Artificial Intelligence*. Harvester Press. Reprinted in [McCarthy, 1990].

[McCarthy, 1979b] McCarthy, J. (1979b). First order theories of individual concepts and propositions. In Michie, D., editor, *Machine Intelligence*, volume 9. Edinburgh University Press, Edinburgh. Reprinted in [McCarthy, 1990].

[McCarthy, 1980] McCarthy, J. (1980). Circumscription—a form of non-monotonic reasoning. *Artificial Intelligence*, 13:27–39. Reprinted in [McCarthy, 1990].

[McCarthy, 1986] McCarthy, J. (1986). Applications of circumscription to formalizing common sense knowledge. *Artificial Intelligence*, 28:89–116. Reprinted in [McCarthy, 1990].

[McCarthy, 1990] McCarthy, J. (1990). *Formalizing Common Sense: Papers by John McCarthy*. Ablex Publishing Corporation, 355 Chestnut Street, Norwood, NJ 07648.

[McCarthy, 1993] McCarthy, J. (1993). Notes on formalizing context. In *IJCAI-93*. Available on http://www-formal.stanford.edu/jmc/.

[McCarthy, 1995a] McCarthy, J. (1995a). Review of "shadows of the mind" by roger penrose. *Psyche*.

[McCarthy, 1995b] McCarthy, J. (1995b). Situation calculus with concurrent events and narrative. available at URL: http://www-formal.stanford.edu/jmc/ Contents subject to change. Reference will remain.

[McCarthy and Hayes, 1969] McCarthy, J. and Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. In Meltzer, B. and Michie, D., editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press.

[Newell, 1980] Newell, A. (1980). Physical symbol systems. *Cognitive Science*, 4:135–183.

[Penrose, 1994] Penrose, R. (1994). *Shadows of the Mind: A Search for the Missing Science of Consciousness.* Oxford University Press, Oxford.

[Sloman, 1985] Sloman, A. (1985). What enables a machine to understand? In *Proceedings 9th International Joint Conference on AI*, pages 995–1001. Morgan-Kaufman.

[Sloman and Croucher, 1981] Sloman, A. and Croucher, M. (1981). Why robots will have emotions. In *Proceedings 7th International Joint Conference on AI.* Morgan-Kaufman.

[Turing, 1950] Turing, A. (1950). Computing machinery and intelligence. *Mind.*