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 to July 15, 2002

[1]

Abstract

Conscious knowledge and other information is distinguished from unconscious information by being observable, and its observation results in conscious knowledge about it. We call this introspective knowledge.

A robot will need to use introspective knowledge in order to operate in the common sense world and accomplish the tasks humans will give it.

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 domains.

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.

Programs with much introspective consciousness do not yet exist.

[1]This paper is substantially changed from [McCarthy, 1996] which was given at Machine Intelligence 15 in 1995 August held at Oxford University.

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. If the advocates of qualia are right, it looks like robots won't need them to exhibit any behavior exhibited by humans.

Contents

1 Introduction **4**
- 1.1 About Logical AI 4
- 1.2 Ascribing mental qualities to systems 5
- 1.3 Consciousness and introspection 6

2 What Consciousness does a Robot Need? **7**
- 2.1 Easy introspection 7
- 2.2 Serious introspection 8
- 2.3 Understanding and Awareness 13

3 Formalized Self-Knowledge **14**
- 3.1 Mental Situation Calculus 15
- 3.2 Mental events, especially mental actions 17

4 Logical paradoxes, Gödel's theorems, and self-confidence **19**
- 4.1 The paradoxes 20
- 4.2 The incompleteness theorems 22
- 4.3 Iterated self-confidence 22
- 4.4 Relative consistency 23

5 Inferring Non-knowledge **23**
- 5.1 Existence of parameterized sets of models 26
- 5.2 Non-knowledge as failure 27

6 Humans and Robots **28**
- 6.1 A conjecture about human consciousness and its consequences for robots 28
- 6.2 Robots Should Not be Equipped with Human-like Emotions . 29

7 Remarks **32**

8 Acknowledgements **35**

1 Introduction

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

This article discusses consciousness with the methodology of logical AI. [McCarthy, 1989] contains a recent discussion of logical AI. AI systems that don't represent information by sentences can have only limited introspective knowledge.

1.1 About Logical AI

[McCarthy, 1959] 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 ([McCarthy, 1980], [McCarthy, 1986], [Brewka, 1991], [Lifschitz, 1994]), theories of concepts as objects [McCarthy, 1979b] and theories of contexts as objects [McCarthy, 1993], [McCarthy and Buvač, 1998]. [McCarthy, 1959] mentioned self-observation but wasn't specific.

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.

1.2 Ascribing mental qualities to systems

A system, e.g. a robot, can be ascribed beliefs provided sentences expressing these beliefs have the right relation to the system's internal states, inputs and output and the goals we ascribe to it. [Dennett, 1971] and [Dennett, 1978] calls such ascriptions the *intentional stance*. The beliefs need not be explicitly represented in the memory of the system. Also Allen Newell, [Newell, 1980] regarded some information not represented by sentences explicitly present in memory as nevertheless representing sentences or propositions believed by the system. Newell called this the *logic level*. I believe he did not advocate general purpose programs that represent information primarily by sentences.[2] I do.

[McCarthy, 1979a] goes into detail about conditions for ascribing belief and other mental qualities.

> 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.

[2] Newell, together with Herbert Simon and other collaborators used logic as a domain for AI in the 1950s. Here the AI was in programs for making proofs and not in the information represented in the logical sentences.

[McCarthy, 1979a] considers systems with very limited beliefs. For example, a thermostat may usefully be ascribed one of exactly three beliefs—that the room is too cold, that it is too warm or that its temperature is ok. This is sometimes worth doing even though the thermostat may be completely understood as a physical system.

Tom Costello pointed out to me that a simple system that doesn't use sentences can sometimes be ascribed some introspective knowledge. Namely, an electronic alarm clock getting power after being without power can be said to know that it doesn't know the time. It asks to be reset by blinking its display. The usual alarm clock can be understood just as well by the design stance as by the intentional stance. However, we can imagine an alarm clock that had an interesting strategy for getting the time after the end of a power failure. In that case, the ascription of knowledge of non-knowledge might be the best way of understanding that part of the state.

1.3 Consciousness and introspection

We propose to design robot *consciousness* with explicitly represented beliefs as follows. At any time a certain set of sentences are directly available for reasoning. We call these the robot's *awareness*. Some of them, perhaps all, are available for observation, i.e. processes can generate sentences about these sentences. These sentences constitute the robot's *consciousness*. In this article, we shall consider the awareness and the consciousness to coincide; it makes the discussion shorter.

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* and constitute *self-consciousness*.

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. [Kraus et al., 1991] has 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. [3]

We discuss what forms of consciousness and introspection are required for robots 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.

2 What Consciousness does a Robot Need?

2.1 Easy introspection

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 finitized version of Turing's argument that the *halting problem* is

[3] Here's an ancient example of observing one's likes and not knowing the reason.

> "Non amo te, Zabidi, nec possum dicere quare;
> Hoc tantum possum dicere, non amo te."

by Martial which Tom Brown translated to

> I do not like thee, Dr. Fell
> The reason why I cannot tell,
> But this I know, I know full well,
> I do not like thee, Dr. Fell.

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.

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.

Very likely, computer programs can be made to get more from reading itself than we presently know how to implement.

The dual concept to programs reading themselves is that of programs modifying themselves. Before the invention of index registers (B-lines) at Manchester, programs did indexing through arrays and telling subroutines where to return by program modification. It was sometimes stated that self-modification was one of the essential ideas of using the same memory for programs and data. This idea went out of fashion when major computers, e.g. the IBM 704 in 1955, had index registers.

As AI advances, programs that modify themselves in substantial ways will become common. However, I don't treat self-modification in this article.

Unfortunately, these easy forms of introspection are insufficient for intelligent behavior in many common sense information situations.

2.2 Serious introspection

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.

This section uses two formalisms described in previous papers.

The first is the notion of a context as a *first class object* introduced in [McCarthy, 1987] and developed in [McCarthy, 1993] and [McCarthy and Buvač, 1998]. As first class objects, contexts can be the values of variables and arguments and values of functions. The most important expression is $Ist(c, p)$, which asserts that the proposition p is true in the context c. Propositions true in *subcontexts* need not be true in *outer contexts*. The language of a subcontext can also be an abbreviated version of the language of an outer context, because the subcontext can involve some assumptions not true in outer contexts. A reasoning system can *enter* a subcontext and reason with the assumptions and in the language of the subcontext. If we have $Ist(c, p)$ in an outer context $c0$, we can write

$$c: \qquad p,$$

and reason directly with the sentence p. Much human reasoning, maybe all, is done in subcontexts, and robots will have to do the same. There is no most general context. The outermost context used so far can always be *transcended* to a yet outer context. A sentence $Ist(c, p)$ represents a kind of introspection all by itself.

The second important formalism is that of a *proposition* or *individual concept* as a first class object distinct from the truth value of the proposition or the value of the individual concept. This allows propositions and individual concepts to be discussed formally in logical language rather than just informally in natural language. One motivating example from [McCarthy, 1979b] is given by the sentences

$$\begin{array}{l} denotation(Telephone(Person)) = telephone(denotation(Person)) \\ denotation(Mike) = mike \\ telephone(mike) = telephone(mary) \\ knows(pat, Telephone(Mike)) \\ \neg knows(pat, Telephone(Mary)). \end{array} \tag{1}$$

Making the distinction between concepts and their denotation allows us to say that Pat knows Mike's telephone number but doesn't know Mary's telephone number even though Mary's telephone number is the same as Mike's telephone number. [McCarthy, 1979b] uses capitalized words for concepts and lower case for objects. This is contrary to the convention in the rest of this paper that capitalizes constants and uses lower case for variables.

We will give tentative formulas for some of the results of observations. In this we take advantage of the ideas of [McCarthy, 1993] and [McCarthy and Buvač, 1998] 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. Already a notebook computer is aware of the state of its battery.

$$\ldots : C(Here, Now, I) : Lowbattery \wedge In(Screwdriver, Hand3) \quad (2)$$

  [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. However, observing that it doesn't know the telephone number and cannot infer what it is involves getting around Gödel's second incompleteness theorem. The reason we have to get around it is that showing that any sentence is not inferrable says that the theory is consistent, because if the theory is inconsistent, all sentences are inferrable. Section 5 shows how do this using Gödel's idea of relative consistency. Consider

$$C(Now, I) : \neg Know(Telephone(Clinton)) \quad (3)$$

  and

$$C(Now, I) : \neg Know\text{-}whether(Sitting(Clinton)). \quad (4)$$

  Here, as discussed in [McCarthy, 1979b], $Telephone(Clinton)$ stands for the *concept* of Clinton's telephone number, and $Sitting(Clinton)$ is the *proposition* that Clinton is sitting.

  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.

- The robot needs more than just the ability to observe that it doesn't know whether a particular sentence is true. It needs to be able to observe that it doesn't know anything about a certain subject, i.e. that anything about the subject is possible. Thus it needs to be able to say that the members of Clinton's cabinet may be in an arbitrary configuration of sitting and standing. This is discussed in Section 5.1.

- Reasoning about its abilities. "I think I can figure out how to do this". "I don't know how to do that."

- 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. This lets it notice when a subgoal has become irrelevant to a larger goal and then abandon it.

- 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.

- It may also be able to say, "I can tell you how I solved that problem" in a way that takes into account its mental search processes and not just its external actions.

- The obverse of a goal is a constraint. Maybe we will 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.

- 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. The restriction to *up to the present* avoids the paradoxes of self-reference and still preserves the useful generality.

- Knowing what goals it can currently achieve and what its choices are for action. [McCarthy and Hayes, 1969a] showed how a robot could think about its own "free will" by considering the effects of the actions it might take, not taking into account its own internal processes that decide on which action to take.

- A simple (and basic) form of free will is illustrated in the situation calculus formula that asserts that John will do the action that John thinks results in the better situation for him.

$$
\begin{aligned}
&Occurs(Does(John,\\
&\quad \textbf{if}\\
&\text{Thinks-better}(John, Result(Does(John, a1), s), Result(Does(John, a2), s))\\
&\quad \textbf{then}\ a1\\
&\quad \textbf{else}\ a2\\
&\qquad), s).
\end{aligned}
\tag{5}
$$

Here Thinks-better$(John, s1, s2)$ is to be understood as asserting that John thinks $s1$ is better for him than $s2$.

- Besides specific information about its mental state, a robot will need general facts about mental processes, so it can plan its intellectual life.
- 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.
- Probably we can design robots to keep their goals in order so that they won't ever have to say, "I wish I didn't want to smoke."

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.

2.3 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. This includes sentences about how fast or how long it can swim.

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.

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

[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 *understood* women's gymnastics well enough to have coached a world champion.

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. This is discussed in the draft [McCarthy, 1995] and further in [McCarthy and Costello, 1998].

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, 1963] and [McCarthy and Hayes, 1969b], 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 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.

[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.

- Among the propositions that can hold are $Know(p)$ and $Believe(p)$, where p again denotes a proposition. Thus we can have

$$Holds(Know(p), s). \tag{6}$$

- As we will shortly see, sentences like

$$Holds(Know(Not\ Know(p), s) \tag{7}$$

are often useful. The sentence(7) 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. 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 $Telephone(Mike)$.

  The convention used in this section of *telephone* and $Telephone$ is different from the convention in the rest of the article of using capital letters to begin constants (whether individual, functional or predicate constants) and using symbols in lower case letters to denote variables.

  This enables us to say

$$Holds(Knows(Telephone(Mike)), s) \tag{8}$$

to assert knowledge of Mike's telephone number and

$$Holds(Know(Not(Knows(Telephone(Mike)))), s) \tag{9}$$

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 (9) in consciousness might stimulate the robot to look in the phone book.

3.2 Mental events, especially mental actions

Mental events change the situation just as do physical events.

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

- In the simplest formalisms mental events occur sequentially. This corresponds ot a *stream of consciousness*. Whether or not the idea describes human consciousness, it is a design option for robot consciousness.

- $Learn(p)$. The robot learns the fact p. An obvious consequence is

$$Holds(Know(p), Result(Learn(p), s)) \tag{10}$$

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

$$Occurs(Learn(p), s) \rightarrow Holds(F\ Know(p), s), \tag{11}$$

  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

$$Holds(F(p), s) \rightarrow Holds(p, Next(p, s)). \tag{12}$$

[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.

Here $Next(p, s)$ denotes the next situation following s in which p holds. (12) asserts that if $F(p)$ holds in s, then there is a next situation in which p holds. (This $Next$ is not the $Next$ operator used in some temporal logic formalisms.)

- 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

$$Occurs(Forget(p), s) \rightarrow Holds(F(Not(Know(p))), s) \qquad (13)$$

  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

$$Occurs(foo, s) \rightarrow Occurs(Forget(p), s) \qquad (14)$$

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

$$Occurs(\textit{Decide-to-do}\ a, s) \rightarrow Holds(\textit{Intend-to-do}\ a, s). \qquad (15)$$

  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 rather involves *entering a context* $Assume(c, p)$ in which p holds. This formalism is described in [McCarthy, 1993] and [McCarthy and Buvač, 1998].

- 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

$$Occurs(See\ o, s) \rightarrow Holds(F\ Knows\ Did\ See\ o, s). \tag{16}$$

  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 [Reiter, 1993] and his colleagues have made important progress.

4 Logical paradoxes, Gödel's theorems, and self-confidence

You can't always get what you want,
But you can sometimes get what you need.
— Rolling Stones

Logical discoveries, mainly of the 20th century, impose limitations on the formalisms we can used without paradox. Other discoveries place limitations on what can be computed. In essence, the limitations apply to both people and machines, and intelligence can live within the limitations.

4.1 The paradoxes

It has precursors, but Russell's paradox of 1901 shows that the obvious set theory, as proposed by Frege has to be modified in unpleasant ways. Frege's basic idea is to let us define the set of all objects having a given property, in more modern notation

$$\{x|\mathcal{P}(x)\},$$

giving the set of all x with the property $\mathcal{P}$. Thus the set of all red dogs is denoted by $\{x|dog(x) \wedge red(x)\}$, or if the set of dogs is denoted *dogs* and the set of red objects as *reds*, we can also write $\{x|x \in dogs \wedge x \in reds\}$. This notation for forming sets is very convenient and is much used in mathematics. The principle is called *comprehension.*

Bertrand Russell in his 1901 letter to Gottlob Frege pointed out that forming the set

$$rp = \{x|\neg(x \in x)\},$$

i.e. the set of all sets that are not members of themselves, leads promptly to a contradiction. We get $rp \in rp \equiv \neg rp \in rp$.

There are many ways of restricting set theory to avoid the contradiction. The most commonly chosen is that of Zermelo, whose set theory Z allowed only writing $\{x \in A|\mathcal{P}(x)\}$, where A is a previously defined set. This turned out to be not quite enough to represent mathematics and Fraenkel introduce a further axiom schema of *replacement* giving a system now called ZF.

ZF is less convenient than Frege's inconsistent system because of the need to find the set A, and the unrestricted comprehension schema is often used when it is clear that the needed A could be found. [7]

[7] For AI it might be convenient to use unrestricted comprehension as a default, with the default to the limited later by finding an A if necessary. This idea has not been explored yet.

A more direct inconvenience for giving robots consciousness is the paradox discovered by Richard Montague [Montague, 1963] concerning a set of desirable axioms for knowledge of sentences.

We might denote by $knows(person, sentence)$ the assertion that $person$ knows *sentence* and consider this as holding at some time t in in some situation s. However, Montague's paradox arises even when there is only one knower, and we write Kp for the knower knowing the sentence p. Montague's paradoxes arise under the assumption that the language of the sentences p is rich enough for "elementary syntax", i.e. allows quantifiers and operations on sentences or on Gödel numbers standing for sentences.

The axioms are

$$Kp \rightarrow p, \tag{17}$$

$$Kp \rightarrow KKp, \tag{18}$$

and

$$K(Kp \wedge K(p \rightarrow q) \rightarrow Kq). \tag{19}$$

Intuitively these axioms state that if you know something, it's true, if you know something, you know you know it, and you can do modus ponens. Added to this are schemas saying that you know some sentences of elementary logic.

From these, Montague constructed a version of the paradox of the liar. Hence they must be weakened, and there are many weakenings that restore consistency. Montague preferred to leave out elementary syntax, thus getting a form of modal logic.

I think it might be better to weaken (18) by introducing a hierarchy of *introspective knowledge operators* on the idea that knowing that you know something is knowledge at an introspective level.

Suppose that we regard knowledge as a function of time or of the situation. We can slither out of Montague's paradox by changing the axiom $Kp \rightarrow KKp$ to say that if you knew something in the past, you now know that you knew it. This spoils Montague's recursive construction of the paradox.

None of this has yet been worked out for an AI system.

4.2 The incompleteness theorems

Gödel's first incompleteness theorem shows that any consistent logical theory expressive enough for elementary arithmetic, i.e. with addition, multiplication and quantifiers could express true sentences unprovable in the theory.

Gödel's second incompleteness theorem tells that the consistency of the system is one of these unprovable sentences.

The basis of Gödel's proof was the fact that the syntactic computations involved in combining formulas and verifying that a sequence of formulas is a proof can be imitated by arithmetic computations on "Gödel numbers" of formulas. If we have axioms for symbolic computations, e.g. for Lisp computations, then the proofs of Gödel's theorems become much shorter. Shankar [Shankar, 1986] has demonstrated this using the Boyer-Moore prover.

Among the unprovable true sentences is the statement of the theory's own consistency. We can interpret this as saying that the theory lacks self-confidence. Turing, in his PhD thesis, studied what happens if we add to a theory T the statement $consis(T)$ asserting that T is consistent, getting a stronger theory T'. While the new theory has $consis(T)$ as a theorem, it doesn't have $consis(T')$ as a theorem—provided it is consistent. The process can be iterated, and the union of all these theories is $consis^{\omega}(T)$. Indeed the process can again be iterated, as Turing showed, to any constructive ordinal number.

4.3 Iterated self-confidence

Gödel's second incompleteness theorem [Gödel, 1965] tells us that a consistent logical theory $T0$ strong enough to do Peano arithmetic cannot admit a proof of its own consistency. However, if we believe the theory $T0$, we will believe that it is consistent. We can add the statement $consis(T0)$ asserting that $T0$ is consistent to $T0$ getting a stronger theory $T1$. By the incompleteness theorem, $T1$ cannot admit a proof of $consis(T1)$, and so on. Adding consistency statement for what we already believe is a *self–confidence prinicple.*

Alan Turing [Turing, 1939] studied iterated statements of consistency, pointing out that we can continue the iteration of self-confidence to form

$T\omega$, which asserts that all the Tn are consistent. Moreover, the iteration can be continued through the *recursive ordinal numbers*. Solomon Feferman [Feferman, 1962] studied a more powerful iteration principle than Turing's called *transfinite progressions of theories*.

There is no single computable iterative self-confidence process that gets everything. If there were, we could put it in a single logical system, and Gödel's theorem would apply to it.

For AI purposes, $T1$, which is equivalent to induction up to the ordinal ϵ_0 may suffice.

The relevance to AI of Feferman's transfinite progressions is at least to refute naive arguments based on the incompleteness theorem that AI is impossible.

A robot thinking about self-confidence principles is performing a kind of introspection. For this it needs not only the iterates of $T0$ but to be able to think about theories in general, i.e. to use a formalism with variables ranging over theories.

4.4 Relative consistency

When we cannot prove a theory consistent, we can often show that it is consistent provided some other theory, e.g. Peano arithmetic or ZF is consistent.

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. Paul Cohen proved in 1963 that if set theory has any models it has models in which the axiom of choice and the continuum hypothesis are false.

5 Inferring Non-knowledge

[This section and the next have a lot of redundancy. This will be fixed.]

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. [8] 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. [9]

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

[8] 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.

[9] 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 she 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.

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 difficult, 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.*[10]

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

$$(\exists P'_1, \ldots, P'_n\ sits')A(P'_1, \ldots, P'_n, sits') \tag{20}$$

expresses the consistency of the theory, and the sentence

$$(\exists P'_1, \ldots, P'_n\ sits')(A(P'_1, \ldots, P'_n, sits') \wedge \neg sits'(Clinton, s)) \tag{21}$$

expresses the consistency of the theory with the added assertion that Clinton is not sitting in the situation s. [In the above, we use upper case of the predicate constant $Sits$ and lower case for the variable $sits'$.

Then

$$(20) \rightarrow (21) \tag{22}$$

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

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

$$(\exists P_2' P_3')A(P_1, P_2', sits') \wedge sits'(Clinton, s). \tag{23}$$

The straightforward way of proving (23) is to find substitutions for the predicate variables P_2' and $sits'$ that make the matrix of (23) 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

$$sits' = (\lambda x\ ss)(\neg(x = Clinton \wedge ss = s) \vee Sits(x, ss))), \tag{24}$$

and (23) 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.

5.1 Existence of parameterized sets of models

Relative consistency provides a reasonable way of handling single cases of non-knowledge. However, we may want more. For example, suppose we want to say that we know nothing about whether any member of Clinton's cabinet is standing or sitting except (for example) that none of them sits when Clinton is standing in the same room.

The theory should then have lots of models, and we can parameterize them by a set of the standees that is arbitary except for the above condition. Here's a formula using non-knowledge.

$$
\begin{aligned}
&(\forall f)(f \in \{t, f\}^{\text{Clinton-cabinet}} \\
&\rightarrow (\forall x)(x \in \text{Clinton-cabinet} \\
&\qquad \rightarrow \neg Know(Sits(x) \equiv f(x) = t)))
\end{aligned} \tag{25}
$$

but this only tells us that for each member of the cabinet, we don't know whether he is sitting.

We want the stronger formula

$$
\begin{aligned}
&(\forall f)(f \in \{t, f\}^{\text{Clinton-cabinet}} \\
&\neg Know(\neg(\forall x)(x \in \text{Clinton-cabinet} \\
&\qquad Sits(x) \equiv f(x) = t)))
\end{aligned} \tag{26}
$$

which asserts that for all we know, Clinton's cabinet could be standing or sitting in an arbitrary pattern. Here we have had to take a quantifier inside the *Know* function. [McCarthy, 1979b] discusses difficulties in formalizing this and doesn't offer a satisfactory solution.

[McCarthy, 1999] gives a simple way of parameterizing the set of models of a propositional sentence. However, there can be no neat way of parameterizing the models of an arbitary first order theory. Thus parameterizing the set of axioms for group theory would amount to parameterizing the set of all groups, and group theory tells us that there is no straightforward parameterization.

5.2 Non-knowledge as failure

A system based on Horn clauses, e.g. a Prolog program, may treat non-knowledged as failure. Thus if both an attempt to prove Clinton to be sitting and an attempt to prove him standing fail, the system can infer that it doesn't know whether he is sitting or standing. This is likely to be easier than establishing that it is possible that he is standing and possible that he is sitting by finding models.

6 Humans and Robots

Human consciousness is undoubtedly more complicated than the design we propose for robots, but it isn't necessarily better.

The main complication I see is that human self observation, like human vision, is spotty. I pursue the analogy, because much more is accssible to observation and experiment with vision than with self observation.

Subjectively a person feels that he has a visual field with everything in the field accessible with approximately equal resolution. We also feel that colors are associated with points in the visual field. In fact, a person has a blind spot, resolution is much better in the small fovea than elsewhere, the perceived color of an object in the field has no simple relation to the light striking a corresponding point on the retina.

All this is because nature has evolved a vision system that finds out as much as possible about the world with very limited apparatus. For example, the usual objects have colors that can be recognized under varied lighting conditions as being the same color.

We have much less ability to observe human consciousness. However, it would be too good to be true if it consisted of a definite set of observable sentences.

6.1 A conjecture about human consciousness and its consequences for robots

There is a large difference between the human mind and the ape mind, and human intelligence evolved from ape-like intelligence in a short time as evolution goes. Our conjecture is that besides the larger brain, there is one qualitative difference—consciousness. The evolutionary step consisted of making more of the brain state itself observable than was possible for our ape-like ancestors. The consequence was that we could learn procedures that take into account the state of the brain, e.g. previous observations, knowledge or lack of it, etc.

The consequence for AI is that maybe introspection can be introduced into problem solving in a rather simple way—letting actions depend on the state of the mind and not just on the state of the external world as revealed by observation.

This suggests designing logical robots with observation as a subconscious process, i.e. mainly taking place in the background rather than as a result of decisions. Observation results in sentences in consciousness. Deliberate observations should also be possible. The mental state would then be one aspect of the world that is subconsciously observed.

We propose to use contexts as formal objects for robot context, whereas context is mainly subconscious in humans. Perhaps robots should also deal with contexts at least partly subconsciously. I'd bet against it now.

[Much more to come when I get it clear.]
2002 July: It's still not sufficiently clear.

6.2 Robots Should Not be Equipped with Human-like Emotions

Human emotional and motivational structure is likely to be much farther from what we want to design than is human consciousness from robot consciousness.[11]

Some authors, [Sloman and Croucher, 1981], have argued that sufficiently intelligent robots would automatically have emotions somewhat like those of humans. However, I think that it would be possible to make robots with human-like emotions, but it would require a special effort distinct from that required to make intelligent robots. 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.

[11] Cindy Mason in her Emotional Machines home page (http://www.emotionalmachines.com/) expresses a different point of view.

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. [12]

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. According to these notions, paranoia, schizophrenia, depression and other mental illnesses would involve malfunctions of the chemical mechanisms that gate ideas into consciousness. A paranoid who believes the CIA is following him and influencing him with radio waves 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.

6. Hormones analogous to neurostransmitters open synaptic gates to admit whole classes of beliefs into consciousness. They are analogs of similar substances and gates in animals.

7. 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.

8. Building the analog of a chemically influenced gating mechanism would require a special effort.

These facts suggest the following design considerations.

1. We don't want robots to bring ideas into consciousness in an uncontrolled way. Robots that are to react against people (say) considered

[12] 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]

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, e.g. panic, 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, such human-like emotional structures are not an automatic byproduct of human-level intelligence.

3. Another aspect of the human mind that we shouldn't build into robots is that subgoals, e.g. ideas of good and bad learned to please parents, can become independent of the larger goal that motivated them. Robots should not let subgoals come to dominate the larger goals that gave rise to them.

4. 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.

[13]

[13]2001: The Steven Spielberg movie, *Artificial Intelligence* illustrates dangers of making robots that partly imitate humans and inserting them into society. I say "illustrates" rather "than provides evidence for", because a movie can illustrate any proposition the makers want, unrestricted by science or human psychology. In the movie, a robot boy is created to replace a lost child. However, the robot does not grow and is immortal and therefore cannot fit into a human family, although they depict it as programmed to love the bereaved mother. It has additional gratuitous differences from humans.

The movie also illustrates Spielberg's doctrines about environmental disaster and human prejudice against those who are different.

7 Remarks

1. In [Nagel, 1974], Thomas Nagel wrote *"Perhaps anything complex enough to behave like a person would have experiences. But that, if true, is a fact that cannot be discovered merely by analyzing the concept of experience."*. This article supports Nagel's conjecture, both in showing that complex behavior requires something like conscious experience, and in that discovering it requires more than analyzing the concept of experience.

2. 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.

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.

   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 partly 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. John H. Flavell, [Flavell and O'Donnell, 1999] and [John H. Flavell and Flavell, 2000], and his colleagues describe experiments concerning the introspective abilities of people ranging from 3 years old to adulthood. Even 3 year

olds have some limited introspective abilities, and the ability to report on their own thoughts and infer the thoughts of others grows with age. Flavell, et. al. reference other work in this area. This is apparently a newly respectable area of experimental psychology, since the earliest references are from the late 1980s.

6. 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 logical artificial intelligence approach has a good chance of achieving human-level intelligence sooner. They won't tell as much about human intelligence, however.

7. 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.

8. 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 and quite similar.. 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.

9. 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

$$Holds(Hope(p), s) \tag{27}$$

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.

10. 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.

11. 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. However, it may not be so simple if different subgoals each gives rise to wishes, e.g. that the other subgoals were different.

12. 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.

13. Psychologists and philosophers from Aristotle on 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.

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

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

8 Acknowledgements

This work was partly supported by ARPA (ONR) grant N00014-94-1-0775 and partly done in 1994 while the author was Meyerhoff Visiting Professor at the Weizmann Institute of Science, Rehovot, Israel.

More recently, this research has been partly supported by ARPA contract no. USC 621915, the ARPA/Rome Laboratory planning initiative under grant (ONR) N00014-94-1-0775 and ARPA/AFOSR under (AFOSR) grant # F49620-97-1-0207.

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

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

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.

[Dennett, 1971] Dennett, D. C. (1971). Intentional systems. *The Journal of Philosophy*, 68(4):87–106.

[Feferman, 1962] Feferman, S. (1962). Transfinite recursive progressions of axiomatic theories. *J. Symbolic Logic*, 27:259–316.

[Flavell and O'Donnell, 1999] Flavell, J. H. and O'Donnell, A. K. (1999). Development of intuitions about mental experiences. *Enfance*. in press.

[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.

[Gödel, 1965] Gödel, K. (1965). On undecidable propositions of formal mathematical systems. In Davis, M., editor, *The Undecidable*. Raven Press. This is the famous 1931 paper.

[John H. Flavell and Flavell, 2000] John H. Flavell, F. L. G. and Flavell, E. R. (2000). Development of children's awareness of their own thoughts. *Journal of Cognition and Development*, 1:97–112.

[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*. Oxford University Press.

[McCarthy, 1959] McCarthy, J. (1959). Programs with Common Sense[14]. In *Mechanisation of Thought Processes, Proceedings of the Symposium of the National Physics Laboratory*, pages 77–84, London, U.K. Her Majesty's Stationery Office. Reprinted in [McCarthy, 1990].

[McCarthy, 1963] McCarthy, J. (1963). A Basis for a Mathematical Theory of Computation[15]. In Braffort, P. and Hirschberg, D., editors, *Computer*

[14]http://www-formal.stanford.edu/jmc/mcc59.html
[15]http://www-formal.stanford.edu/jmc/basis.html

Programming and Formal Systems, pages 33–70. North-Holland, Amsterdam.

[McCarthy, 1979a] McCarthy, J. (1979a). Ascribing mental qualities to machines[16]. 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[17]. 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[18]. *Artificial Intelligence*, 13:27–39. Reprinted in [McCarthy, 1990].

[McCarthy, 1986] McCarthy, J. (1986). Applications of Circumscription to Formalizing Common Sense Knowledge[19]. *Artificial Intelligence*, 28:89–116. Reprinted in [McCarthy, 1990].

[McCarthy, 1987] McCarthy, J. (1987). Generality in artificial intelligence. *Communications of the Association for Computing Machinery*, 30:1030–1035. Reprinted in [McCarthy, 1990].

[McCarthy, 1989] McCarthy, J. (1989). Artificial Intelligence, Logic and Formalizing Common Sense[20]. In Thomason, R., editor, *Philosophical Logic and Artificial Intelligence*. Klüver Academic.

[McCarthy, 1990] McCarthy, J. (1990). *Formalizing Common Sense: Papers by John McCarthy*. Ablex Publishing Corporation.

[16]http://www-formal.stanford.edu/jmc/ascribing.html
[17]http://www-formal.stanford.edu/jmc/concepts.html
[18]http://www-formal.stanford.edu/jmc/circumscription.html
[19]http://www-formal.stanford.edu/jmc/applications.html
[20]http://www-formal.stanford.edu/jmc/ailogic.html

[McCarthy, 1993] McCarthy, J. (1993). Notes on Formalizing Context[21]. In *IJCAI-93*.

[McCarthy, 1995] McCarthy, J. (1995). Situation Calculus with Concurrent Events and Narrative[22]. Web only, partly superseded by [McCarthy and Costello, 1998].

[McCarthy, 1996] McCarthy, J. (1996). Making Robots Conscious of their Mental States[23]. In Muggleton, S., editor, *Machine Intelligence 15*. Oxford University Press. Appeared in 2000. The web version is improved from that presented at Machine Intelligence 15 in 1995.

[McCarthy, 1999] McCarthy, J. (1999). Parameterizing models of propositional calculus formulas[24]. *web only for now*.

[McCarthy and Buvač, 1998] McCarthy, J. and Buvač, S. (1998). Formalizing Context (Expanded Notes). In Aliseda, A., Glabbeek, R. v., and Westerståhl, D., editors, *Computing Natural Language*, volume 81 of *CSLI Lecture Notes*, pages 13–50. Center for the Study of Language and Information, Stanford University.

[McCarthy and Costello, 1998] McCarthy, J. and Costello, T. (1998). Combining narratives. In *Proceedings of Sixth Intl. Conference on Principles of Knowledge Representation and Reasoning*, pages 48–59. Morgan-Kaufman.

[McCarthy and Hayes, 1969a] McCarthy, J. and Hayes, P. J. (1969a). Some Philosophical Problems from the Standpoint of Artificial Intelligence[25]. In Meltzer, B. and Michie, D., editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press. Reprinted in [McCarthy, 1990].

[McCarthy and Hayes, 1969b] McCarthy, J. and Hayes, P. J. (1969b). Some Philosophical Problems from the Standpoint of Artificial Intelligence[26]. In

[21]http://www-formal.stanford.edu/jmc/context.html
[22]http://www-formal.stanford.edu/jmc/narrative.html
[23]http://www-formal.stanford.edu/jmc/consciousness.html
[24]http://www-formal.stanford.edu/jmc/parameterize.html
[25]http://www-formal.stanford.edu/jmc/mcchay69.html
[26]http://www-formal.stanford.edu/jmc/mcchay69.html

Meltzer, B. and Michie, D., editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press. Reprinted in [McCarthy, 1990].

[Montague, 1963] Montague, R. (1963). Syntactical treatments of modality, with corollaries on reflexion principles and finite axiomatizability. *Acta Philosophica Fennica*, 16:153–167. Reprinted in [Montague, 1974].

[Montague, 1974] Montague, R. (1974). *Formal Philosophy*. Yale University Press.

[Nagel, 1974] Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, 83(4):435–50.

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

[Reiter, 1993] Reiter, R. (1993). Proving properties of states in the situation calculus. *Artificial Intelligence*, 64:337–351. available from http://www.cs.utoronto.ca/ cogrobo.

[Shankar, 1986] Shankar, N. (1986). *Proof-Checking Metamathematics*. PhD thesis, Computer Science Department, University of Texas at Austin.

[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*.

[Turing, 1939] Turing, A. M. (1939). Systems of logic based on ordinals. *Proc Lond Math Soc (2)*, 45.