NOTES ON SELF-AWARENESS

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

2004 Apr 11, 10:41 p.m.

Abstract

These notes discuss self-awareness in humans and machines. The goal is to determine useful forms of machine self-awareness and also those that are on the road to human-level AI.

This is a draft which is to be improved, and suggestions are solicited. There are a few formulas in this version. The final version will have more.

1 Introduction

Developing self-aware computer systems will be an interesting and challenging project. It seems to me that the human forms of self-awareness play an important role in humans achieving our goals and will also be important for advanced computer systems. However, I think they will be difficult to implement in present computer formalisms, even in the most advanced logical AI formalisms. The useful forms of computer agent self-awareness will not be identical with the human forms. Indeed many aspects of human self-awareness are bugs and will not be wanted in computer systems. (McCarthy 1996) includes a discussion of this and other aspects of robot consciousness.

Nevertheless, for now, human self-awareness, as observed introspectively is the best clue. Introspection may be more useful than the literature of experimental psychology, because it gives more ideas, and the ideas can be checked for usefulnes by considering programs that implement them. Moreover, at least in the beginning of the study of self-awareness, we should be ontologically promiscuous, e.g. we should not identify intentions with goals Significant differences may become apparent, and we can always squeeze later. [1]

Some human forms of self-awareness are conveniently and often linguistically expressed and others are not. For example, one rarely has occasion to announce the state of tension in ones muscles. However, something about it can be expressed if useful. How the sensation of blue differs from the sensation of red apparently cannot be verbally expressed. At least the qualia-oriented philosophers have put a lot of effort into saying so. What an artificial agent can usefully express in formulas need not correspond to what humans ordinarily say, or even can say. In general, computer programs can usefully be given much greater powers of self-awareness than humans have, because every component of the state of the machine or its memory can be made accessible to be read by the program.

A straightforward way of logically formalizing self-awareness is in terms of a mental situation calculus with certain observable fluents. The agent is aware of the observable mental fluents and their values. A formalism with mental situations and fluents will also have mental events including actions, and their occurrence will affect the values of the observable fluents. I advocate the form of situation calculus proposed in (McCarthy 2002).

Self-awareness is continuous with other forms of awareness. Awareness of being hot and awareness of the room being hot are similar. A simple fluent of which a person is aware is hunger. We can write $Hungry(s)$ about a mental situation s, but we write $Holds(Hungry, s)$, then $Hungry$ can be the value of bound variables.. Anohter advantage is that now $Hungry$ is an object, and the agent can compare $Hungry$ with $Thirsty$ or $Bored$. I'm not sure where the object $Hunger$ comes in, but I'm pretty sure our formalism should have it and not just $Hungry$. We can even use $Holds(Applies(Hunger, I), s)$

[1] Some philosophers who emphasize qualia may be inclined to regard self-awareness as a similar phenomenon—in which a person has an undifferentiated awareness of self, like the qualia oriented notion of the pure sensation of red as distinct from blue. This is not at all what is needed for AI. Rather we study the specific aspects of self and its activity which it is useful to be aware.

but tolerate abbreviations, especially in contexts. [2] [3] [4]

Our goal in this research is an *epistemologically adequate* formalism in the sense of (McCarthy and Hayes 1969) for representing what a person or robot can actually learn about the world. In this case, the goal is to represent facts of self-awareness of a system, both as an internal language for the system and as an external language for the use of people or other systems.

Basic entities, e.g. automaton states as discussed in (McCarthy and Hayes 1969) or neural states may be good for devising theories at present, but we cannot express what a person or robot actually knows about its situation in such terms.

2 Of what are we aware, and of what should computers be aware?

Humans are aware of many different aspects of their minds. Here are samples of kinds of self-awareness—alas not a classification.

[2]The English grammatical form "I am hungry" has no inevitability to it. French has "*J'ai faim*", literally "I have hunger", and German has "*Es hungert mich*", literally "It hungers me". In French the noun "faim" meaning "hunger" is used, whereas in English an adjective "hungry" is used. In logical we have both; I don't see a use for a logical version of the German form.

[3]$Holds(Hungry(I), s)$ might be written if our agent needs to compare its hunger with that of other agents. However, if we use formalized contexts ((McCarthy 1993)) we can get by with $Holds(Hungry, s)$ in an inner context in which the sentences are about the agent's self. We won't use formalized contexts in these notes, but an informal notion of context can avoid some worries. For example, some discussions are carried out entirely in contexts in which the fact that John McCarthy is a professor at Stanford is permanent. However, when needed, this context can be transcended. Likewise there are useful time-limited contexts in which George W. Bush is permanently President of the United States.

[4]In spite of the fact that English has an enormous vocabulary, the same word is used with diverse meanings. I don't speak of simple homonyms like "lock on a door" and "lock of hair". These can be ruthlessly eliminated from our computer language, e.g. by having words lock1 and lock2. A more interesting example is that one can speak of knowing a person, knowing a fact, and knowing a telephone number. German uses *kennen* for the first and *wissen* for the second; I don't know about the third. In my (McCarthy 1979), "First order theories of individual concepts and propositions", I use different words for the different concepts. I suspect that it will be useful to tolerate using the same term in related senses, e.g. using the same word for the bank as an institution and as a building, because too many related meanings will arise.

1. Permanent aspects of self and their relations to each other and aspects of other persons.

   Thus I am human like other humans. [I am a small child, and I am "supposed to do" what the others are doing. This is innate or learned very early on the basis of an innate predisposition to learn it.[5]

   What might we want an artificial agent to know about the fact that it is one among many agents? It seems to me that the forms in which self-awareness develops in babies and children are likely to be particularly suggestive for what we will want to build into computers.

2. I exist in time. This is distinct from having facts about particular time, but what use can we make of the agent knowing this fact—or even how is the fact to be represented?

3. I don't know Spanish but can speak Russian and French a little. Similarly I have other skills.

   It helps to organize as much as possible of a system's knowledge as knowledge of permanent entities.

4. I often think of you. I often have breakfast at Caffe Verona.

5. Ongoing processes

   I am driving to the supermarket. One is aware of the past of the process and also of its future. Awareness of its present depends on some concept of the "extended now".

   Temporary phenomena

6. Wants, intentions and goals:

   Wants can apply to both states and actions. I want to be healthy, wealthy and wise. I want to marry Yumyum and plan to persuade her guardian Koko to let me.

7. I intend to walk home from my office, but if someone offers me a ride, I'll take it. I intend to give X a ride home, but if X doesn't want it, I won't.

[5] Autistic children may be deficient in this respect.

8. If I intend to drive via Pensacola, Florida, I'll think about visiting Pat Hayes.

   I suppose you can still haggle, and regard intentions as goals, but if you do you are likely to end up distinguishing a particular kind of goal corresponding to what the unsophisticated call an intention.

9. Attitudes

   Attitudes towards the future:
   hopes, fears, goals, expectations, anti-expectations, intentions action: predict, want to know, promises and commitments.

   Attitudes toward the past:
   regrets, satisfactions, counterfactuals
   I'm aware that I regret having offended him. I believe that if I hadn't done so, he would have supported my position in this matter. It looks like a belief is a kind of weak awareness.

   Attitudes to the present:
   satisfaction, I see a dog. I don't see the dog. I wonder where the dog has gone.

   There are also attitudes toward timeless entities, e.g. towards kinds of people and things. I like strawberry ice cream but not chocolate chip.

10. Hopes: A person can observe his hopes. I hope it won't rain tomorrow. Yesterday I hoped it wouldn't rain today. I think it will be advantageous to equip robots with mental qualities we can appropriately call hopes.

11. Fears: I fear it will rain tomorrow. Is a fear just the opposite of a hope? Certainly not in humans, because the hormonal physiology is different, but maybe we could design it that way in robots. Maybe, but I wouldn't jump to the conclusion that we should.

    Why are hopes and fears definite mental objects? The human brain is always changing but certain structures can persist. Specific hopes and fears can last for years and can be observed. It is likely to be worthwhile to build such structures into robot minds, because they last much longer than specific neural states.

12. An agent may observe that it has incompatible wants.

2.1 Mental actions

The companion of observation is action. A theory of self-awareness, i.e. of mental observation, is complemented by a theory of mental action.

(McCarthy 1982) discusses heuristics for coloring maps with four colors. A form of self-awareness is involved. In coloring a map of the United States, the goal of coloring California can be postponed to the very end, because it has only three neighbors and therefore no matter how the rest of the map is colored, there will always be a color for California. Once California is removed from the map, Arizona has only three neighbors. The *postponement* process can be continued as long as possible. In the case of the US, all states get postponed and then can be colored without backtracking. In general it is often possible to observe that in planning a task, certain subtasks can be postponed to the end. Thus postponement of goals is a mental action that is sometimes useful.

A human-level agent, and even an agent of considerably lower level, has policies. These are creatable, observable both in their structure and in their actions, and changeable.

Useful actions: decide on an intention or a goal. Drop an intention.

Clearly there are many more mental actions and we need axioms describing their effects.

3 Machine self-awareness

Self-awareness is not likely to be a feasible or useful attribute of a program that just computes an answer. It is more likely to be feasible and useful for programs that maintain a persistent activity.

What kind of program would be analogous to Pat deciding while on the way to his job that he needed cigarettes? See formula (4) below. Here are some possibiliities.

It's not clear what event in a computer program might correspond to Pat's sudden need for cigarettes. The following examples don't quite make it.

A specialized theorem-proving program $T1$ is being operated as a subprogram by a reasoning program T. Assume that the writer of T has only limited knowledge of the details of $T1$, because $T1$ is someone else's program. T might usefully monitor the operation of $T1$ and look at the collection of

intermediate results $T1$ has produced. If too many of these are redundant, T may restart $T1$ with different initial conditions and with a restriction that prevents sentences of a a certain form from being generated.

An operating system keeps track of the resources a user is using, and check for attempts to use forbidden resources. In particular it might check for generation of password candidates. In its present form this example may bad, because we can imagine the checking be done by the programs that implement supervisor calls rather than by an inspector operating with clock interrupts. While the programs called by clock interrupts exhibit a simple form of self-awareness, the applications I know about are all trivial.

The main technical requirement for self-awareness of ongoing processes in computers is an interrupt system, especially a system that allows clock interrupts. Hardware supporting interrupts is standard on all computers today but didn't become standard until the middle 1960s.[6] The human brain is not a computer that executes instructions in sequence and therefore doesn't need an interrupt system that can make it take an instruction out of sequence. However, interruption of some kind is clearly a feature of the brain.

With humans the boundary between self and non-self is pretty clear. It's the skin. With computer based systems, the boundary may be somewhat arbitrary, and this makes distinguishing self-awareness from other awareness arbitrary. I suppose satisfactory distinctions will become clearer with experience.

3.1 Interrupts, programming languages and self-awarenes

Consider a persistent program driving a car that is subject to observation and modification by a higher level program. We mentioned the human example of noticing that cigarettes are wanted and available. The higher level program must observe and modify the state of the driving program. It seems that a

[6]Historical note: The earliest interrupt system I know about was the "Real time package" that IBM announced as a special order instigated by Boeing in the late 1950s. Boeing wanted it in order to control a wind tunnel using an IBM 704 computer. At M.I.T. we also needed an interrupt system in order to experiment with time-sharing on the IBM 704. We designed a simple one, but when we heard about the much better "Real time package" we started begging for it. Begging from IBM took a while, but they gave it to us.

The earliest standard interrupt system was on the D.E.C. PDP-1 and was designed by Ed Fredkin, then at Bolt, Beranek and Newman, who persuaded D.E.C. to include it in the machine design. Again the purpose was time-sharing.

clock interrupt activating the higher level program is all we need from the hardware.

We need considerably more from the software and from the programming languages. A cursory glance at the interrupt handling facilities of C, Ada, Java, and Forth suggests that they are suitable for handling interrupts of high level processes by low level processes that buffer the transfer of information.

Lisp and Smalltalk can handle interrupts, but have no standard facilities.

My opinion, subject to correction, is that self-awareness of the kinds proposed in this note will require higher level programming language facilities whose nature may be presently unknown. They will be implemented by the present machine language facilities.

However, one feature of Lisp, that programs are data, and their abstract syntax is directly represented, is likely to be necessary for programs that examine themselves and their subprograms. This feature of Lisp hasn't been much used except in macros and has been abandoned in more recent programming languages—in my opinion mistakenly.

4 Formulas

Formalized contexts as discussed in (McCarthy 1993) will be helpful in expressing self-awareness facts compactly.

Pat is aware of his intention to eat dinner at home.

$$\begin{array}{l} c(Awareness(Pat)) : Intend(I, Mod(At(Home), Eat(Dinner))) \\ \text{or} \\ Ist(Awareness(Pat), Intend(I, Mod(At(Home), Eat(Dinner)))) \end{array} \tag{1}$$

Here $Awareness(Pat)$ is a certain context. $Eat(Dinner)$ denotes the general act of eating dinner, logically different from eating $Steak7642$. $Mod(At(Home), Eat(Dinner))$ is what you get when you apply the modifier "at home" to the act of eating dinner. I don't have a full writeup of this proposal for handling modifiers like adjectives, adverbs, and modifier clauses. $Intend(I, X)$ says that I intend X. The use of I is appropriate within the context of a person's (here Pat's) awareness.

We should extend this to say that Pat will eat dinner at home unless his intention changes. This can be expressed by formulas like

$$\neg Ab17(Pat, x, s) \wedge Intends(Pat, Does(Pat, x), s) \rightarrow (\exists s' > s) Occurs(Does(Pat, x), s). \tag{2}$$

in the notation of (McCarthy 2002).

Here's an example of awareness leading to action.

Pat is driving to his job. Presumably he could get there without much awareness of that fact, since the drive is habitual. However, he becomes aware that he needs cigarettes and that he can stop at Mac's Smoke Shop and get some. Two aspects of his awareness, the driving and the need for cigarettes are involved. That Pat is driving to his job can be expressed with varying degrees of elaboration. Here are some I have considered.

$$\begin{array}{l} Driving(Pat, Job, s) \\ \\ Doing(Pat, Drive(Job), s) \\ \\ Holds(Doing(Pat, Mod(Destination(Job), Drive)), s) \\ \\ Holds(Mod(Ing, Mod(Destination(Job, Action(Drive, Pat))), s) \\ . \end{array} \tag{3}$$

The last two use a notion like that of an adjective modifying a noun. Here's a simple sentence giving a consequence of Pat's awareness. It uses *Aware* as a modal operator. This may require repair or it may be ok in a suitably defined context.

$$\begin{array}{l} Aware(Pat, Driving(Job, s), s) \wedge Aware(Pat, Needs(Cigarettes), s) \\ \wedge Aware(Pat, About\text{-}to\text{-}pass(CigaretteStore, s), s) \\ \rightarrow Occurs(StopAt(CigaretteStore), s). \end{array} \tag{4}$$

The machine knows that if its battery is low, it will be aware of the fact.

$$Knows(Machine, (\forall s')(LowBattery(s') \rightarrow Aware(LowBattery(s'))), s) \tag{5}$$

The machine knows, perhaps because a sensor is broken, that it will not necessarily be aware of a low battery.

$$Knows(Machine, \neg(\forall s')(LowBattery(s') \rightarrow Aware(LowBattery(s'))), s) \tag{6}$$

The positive sentence "I am aware that I am aware ..." doesn't seem to have much use by itself, but sentences of the form "If X happens, I will be aware of Y" should be quite useful.

5 Miscellaneous

Here are some examples of awareness and considerations concerning awareness that don't yet fit the framework of the previous sections.

I am slow to solve the problem because I waste time thinking about ducks. I'd like Mark Stickel's SNARK to observe, "I'm slow to solve the problem, because I keep proving equivalent lemmas over and over".

I was aware that I was letting my dislike of the man influence me to reject his proposal unfairly.

Here are some general considerations about what fluents should be used in making self-aware systems.

1. Observability. One can observe ones intentions. One cannot observe the state of ones brain at a more basic level. This is an issue of epistemological adequacy as introduced in (McCarthy and Hayes 1969).

2. Duration. Intentions can last for many years, e.g. "I intend to retire to Florida when I'm 65". "I intend to have dinner at home unless something better turns up."

3. Forming a system with other fluents. Thus beliefs lead to other beliefs and eventually actions.

Is there a technical difference between observations that constitute self-observations and those that don't? Do we need a special mechanism for *self-observation*? At present I don't think so.

If p is a precondition for some action, it may not be in consciousnes, but if the action becomes considered, whether p is true will then come into consciousnes, i.e. short term memory. We can say that the agent is *subaware* of p.

What programming languages provide for interrupts?

References

McCarthy, J. 1979. Ascribing mental qualities to machines[7]. In M. Ringle (Ed.), *Philosophical Perspectives in Artificial Intelligence*. Harvester Press. Reprinted in (McCarthy 1990).

McCarthy, J. 1982. Coloring maps and the Kowalski doctrine[8]. Technical Report STAN-CS-82-903, Dept Computer Science, Stanford University, April. AIM-346.

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

McCarthy, J. 1993. Notes on Formalizing Context[9]. In *IJCAI93*.

McCarthy, J. 1996. Making Robots Conscious of their Mental States[10]. In S. Muggleton (Ed.), *Machine Intelligence 15*. Oxford University Press. Appeared in 2000. The web version is improved from that presented at Machine Intelligence 15 in 1995.

McCarthy, J. 2002. Actions and other events in situation calculus[11]. In B. S. A.G. Cohn, F. Giunchiglia (Ed.), *Principles of knowledge representation and reasoning: Proceedings of the eighth international conference (KR2002)*. Morgan-Kaufmann.

McCarthy, J., and P. J. Hayes. 1969. Some Philosophical Problems from the Standpoint of Artificial Intelligence[12]. In B. Meltzer and D. Michie (Eds.), *Machine Intelligence 4*, 463–502. Edinburgh University Press. Reprinted in (McCarthy 1990).

/@steam.stanford.edu:/u/jmc/f03/selfaware.tex: begun Sun Oct 26 12:45:55 2003, latexed April 11, 2004 at 10:41 p.m.

[7] http://www-formal.stanford.edu/jmc/ascribing.html
[8] http://www-formal.stanford.edu/jmc/coloring.html
[9] http://www-formal.stanford.edu/jmc/context.html
[10] http://www-formal.stanford.edu/jmc/consciousness.html
[11] http://www-formal.stanford.edu/jmc/sitcalc.html
[12] http://www-formal.stanford.edu/jmc/mcchay69.html