What Will Self-Aware Computer Systems Be?

John McCarthy, Stanford University
mccarthy@stanford.edu
http://www-formal.stanford.edu/jmc/
September 13, 2004

- Darpa Wants To Know, And There's A Workshop Tomorrow.

- The Subject Is Ready For Basic Research.

- Short Term Applications May Be Feasible.

- Self-Awareness Is Mainly Applicable To Programs With Persistent Existence.

WHAT WILL SELF-AWARE SYSTEMS BE AWARE OF?

- Easy aspects of state: battery level, memory available, etc.
- Ongoing activities: serving users, driving a car
- Knowledge and lack of knowledge
- purposes, intentions, hopes, fears, likes, dislikes
- Actions it is free to choose among relative to external constraints. That's where free will comes from.
- Permanent aspects of mental state, e.g. long term goals, beliefs,
- Episodic memory—only partial in humans, probably absent in animals, but readily available in computer systems.

HUMAN SELF-AWARENESS—1

• Human self-awareness is weak but improves with age.

• Five year old but not three year old. I used to think the box contained candy because of the cover, but now I know it has crayons. He will think it contans candy,

• Simple examples: I'm hungry, my left knee hurts from a scrape, my right knee feels normal, my right hand is making a fist.

• Intentions: I intend to have dinner, I intend to visit New Zealand some day. I do not intend to die.

• I exist in time with a past and a future. Philosophers argue a lot about what this means and how to represent it.

• Permanent aspects of ones mind: I speak English and a little French and Russian. I like hamburgers and caviar. I cannot know my blood pressure without measuring it.

HUMAN SELF-AWARENESS—2

- What are my choices? (Free will is having choices.)

- Habits: I know I often think of you. I often have breakfast at the Pennsula Creamery.

- Ongoing processes: I'm typing slides and also getting hungry.

- Juliet hoped there was enough poison in Romeo's vial to kill her.

- More: fears, wants (sometimes simultaneous but incompatible)

- Permanent compared with instantaneous wants.

MENTAL EVENTS (INCLUDING ACTIONS)

- consider
- Infer
- decide
- ccdhoose to believe
- remember
- forget
- realize
- ignore

MACHINE SELF-AWARENESS

- Easy self-awareness: battery state, memory left

- Straightorward s-a: the program itself, the programming language specs, the machine specs.

- Self-simulation: Any given number of steps, can't do in general "Will I ever stop?", "Will I stop in less than n steps in general—in less than n steps.

- Its choices and their inferred consequences (free will).

- "I hope it won't rain tomorrow". Should a machine hope and be aware that it hopes? I think it should sometimes.

- $\neg Knows(I, TTelephone(MMike))$, so I'll have to look it up.

WHY WE NEED CONCEPTS AS OBJECTS

We had $\neg Knows(I, TTelephone(MMike))$, and I'll have to look it up.

Suppose $Telephone(Mike) =$ "321-7580". If we write

$\neg Knows(I, Telephone(Mike))$, then substitution would give $\neg Knows(I,$ "321-7580"), which doesn't make sense.

There are various proposals for getting around this. The most advocated is some form of modal logic. My proposal is to regard *individual concepts* as objects, and represent them by different symbols, e.g. doubling the first letter.

There's more about why this is a good idea in my "First order theories of individual concepts and propositions"

WE ALSO NEED CONTEXTS AS OBJECTS

We write

$$c : p$$

to assert p while in the context c. Terms also can be written using contexts. $c : e$ is an expression e in the context c.

The main application of contexts as objects is to assert relations between the objects denoted by different expressions in different contexts. Thus we have

$$c : Does(Joe, a) = SpecializeActor(c, Joe) : a,$$

or, more generally,

$$SpecializesActor(c, c', Joe) \rightarrow c : Does(Joe, a)) = c' : a.$$

Such relations between expressions in different contexts allows using a situation calculus theory in which the actor is not explicitly represented in an outer context in which there is more than one actor.

We also need to express the relation between an external context in which we refer to the knowledge and awareness of AutoCar1 and AutoCar1's internal context in which it can use "I".

SELF-AWARENESS EXPRESSED IN LOGICAL FORMULAS—1

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

$$c(Awareness(Pat)) : Intend(I, MMod(AAt(HHome), EEat(DDinner \tag{1}$$

$Awareness(Pat)$ is a context. $Eat(Dinner)$ denotes the general act of eating dinner, logically different from eating $Steak$7642. $Mod(At(Home), Eat(Dinner))$ is what you get when you apply the modifier "at home" to the act of eating dinner. $Intend(I, X)$ 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

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

in the notation of [**?**].

FORMULAS—2

• AutoCar1 is driving John from Office to Home. AutoCar1 is aware of this. Autocar1 becomes aware that it is low on hydrogen. AutoCar1 is permanently aware that it must ask permission to stop for gas, so it asks for permission. Etc., Etc. These facts are expressed in a context $C0$.

$$
\begin{aligned}
&C0: \\
&Driving(I, John, Home1) \\
&\wedge Aware(DDriving(II, JJohn, HHome) \\
&\wedge OccursBecomes(Aware(I, LLowfuel(AAutoCar1))) \\
&\wedge OccursBecomes(Want(I, SStopAt(GGasStation1))) \\
&\wedge
\end{aligned}
\tag{3}
$$

QUESTIONS

• Does the lunar explorer require self-awareness? What about the entries in the recent DARPA contest?

• Do self-aware reasoning systems require dealing with referential opacity? What about explicit contexts?

• Where does tracing and journaling involve self-awareness?

• Does an online tutoring program (for example, a program that teaches a student Chemistry) need to be self aware?

• What is the simplest self-aware system?

- Does self-awareness always involve self-monitoring?

- In what ways does self-awareness differ from awareness of other agents? Does it require special forms of representation or architecture?