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 T

- The Subject Is Ready For Basic Research.

- Short Term Applications May Be Feasible.

- Self-Awareness Is Mainly Applicable To Programs W
tent Existence.

WHAT WILL SELF-AWARE SYSTEMS BE AWA

- Easy aspects of state: battery level, memory availa
- 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 ext straints. That's where free will comes from.
- Permanent aspects of mental state, e.g. long t beliefs,
- Episodic memory—only partial in humans, probably animals, but readily available in computer systems.

HUMAN SELF-AWARENESS—1

• Human self-awareness is weak but improves with ag

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

• Simple examples: I'm hungry, my left knee hurts fro
my right knee feels normal, my right hand is making

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

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

• Permanent aspects of ones mind: I speak English a
French and Russian. I like hamburgers and caviar. I ca
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 b
the Pennsula Creamery.

- Ongoing processes: I'm typing slides and also getti

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

- More: fears, wants (sometimes simultaneous but inc

- 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 prograr guage specs, the machine specs.

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

- Its choices and their inferred consequences (free wi

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

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

WHY WE NEED CONCEPTS AS OBJECT

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

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

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

There are various proposals for getting around this.
advocated is some form of modal logic. My proposal i
individual concepts as objects, and represent them b
symbols, e.g. doubling the first letter.

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

WE ALSO NEED CONTEXTS AS OBJECT

We write

$$c : p$$

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

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

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

or, more generally,

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

Such relations between expressions in different conte
using a situation calculus theory in which the actor is
itly represented in an outer context in which there is
one actor.

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

SELF-AWARENESS EXPRESSED IN LOGIC FORMULAS—1

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

$$c(Awareness(Pat)) : Intend(I, MMod(AAt(HHome), E$$

$Awareness(Pat)$ is a context. $Eat(Dinner)$ denotes t act of eating dinner, logically different from eating St $Mod(At(Home), Eat(Dinner))$ is what you get when the modifier "at home" to the act of eating dinner. I says that I intend X. The use of I is appropriate context of a person's (here Pat's) awareness.

We should extend this to say that Pat will eat dinne
unless his intention changes. This can be expressed b
like

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

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

FORMULAS—2

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

$C0$:
$Driving(I, John, Home1)$
$\wedge Aware(DDriving(II, JJohn, HHome)$
$\wedge OccursBecomes(Aware(I, LLowfuel(AAutoCar1)$
$\wedge OccursBecomes(Want(I, SStopAt(GGasStation1)$
$\wedge$

QUESTIONS

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

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

• Where does tracing and journaling involve self-awar

• Does an online tutoring program (for example, a pro
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 awarene
agents? Does it require special forms of representatio
tecture?