WHAT CONSCIOUSNESS DOES A ROBOT NEED

John McCarthy
Computer Science Department
Stanford University
jmc@cs.stanford.edu
http://www-formal.stanford.edu/jmc/
July 12, 2002

Almost all of my papers are on the web page. This paper is http://www-formal.stanford.edu/consciousness.html
APPROACHES TO ARTIFICIAL INTELLIGENCE

**biological**—Humans are intelligent; imitate humans observe and imitate at either the psychological or neurophysiological level

**engineering**—The world presents problems to intelligence. Study information and action available in the world.
1. Write programs using non-logical representations.
2. Represent facts about the world in logic and decide what to do by logical inference

We aim at human level AI, and the key phenomenon is the common sense informatic situation.
THE COMMON SENSE INFORMATIC SITUATION

• Involves approximate entities.

• There is no limitation on what information may be relevant. Theories must be elaboration tolerant.

• Reasoning must often be non-monotonic.

Common sense theories therefore contrast with formal scientific theories and most present AI theories. Science is embedded in common sense.
A LOGICAL ROAD TO HUMAN LEVEL AI

• Use *Drosophila* that illustrate aspects of representation and reasoning problems.

• Concepts, context, circumscription, counterfactuals, consciousness, creativity, approximation

• narrative, projection, planning

• mental situation calculus

• domain dependent control of reasoning
Logic in AI

Features of the logic approach to AI.

- Represent information by sentences in a logical language, e.g. first order logic, second order logic, modal logic, theory in logic.

- Auxiliary information in tables, programs, states, etc., described by logical sentences.

- Inference is logical inference—deduction supplemented by some form of nonmonotonic inference, e.g. circumscription.
• Action takes place when the system infers that it should do the action.

• Observation of the environment results in sentences in memory.

• Situation calculus formalizes the relations $\text{holds}(p, s)$, $\text{occurs}(e, s)$ and the function $\text{result}(e, s)$ which has a new situation as its value.

• Formalizing consciousness involves giving situations mental components.

• Self-observation results in sentences about the system's state of mind.
What Introspection do Robots Need?

- **What’s this?:** What ability to observe its own computational state and computational processes does a robot need to do its tasks?

- **General Knowledge?:** What general information about reasoning processes does it need to plan its mental life?

- **Design approach:** Asking what consciousness is needed gives different answers from those trying to define consciousness has given.
• **Recommendation for AI**: Introspection is needed to decide whether to think or look, to learn from near misses, use counterfactuals and keep pedigrees of beliefs.

• **Recommendation for psychologists and philosophers**: Add this *direct design stance* approach to your methodology.
What is Consciousness? We consider several kinds of knowledge.

- There are many unconscious stimulus-response relations in animal and humans, and there can be in machines.

- Unconscious knowledge can affect behavior.

- Conscious knowledge and other conscious information can be observed by the actor.

- Self-conscious knowledge is conscious knowledge about conscious information.

- Some aspects of behavior require decisions of the whole system. Which way to run is an example. These decisions are made by a central mechanism.
• In logical robots, the *consciousness* is be a sub-region of memory containing facts and other mental entities.

• Reasoning involves the entities in consciousness and leads to decisions when the reasoning leads to a statement that an action should be performed.

• The capacity of consciousness is limited, so new information displaces old, which may go to a history file.
Taxonomy of Consciousness

• The consciousness itself can be observed and the obser-
vations enter consciousness.

• Robot consciousness can be given powers people do not have.
  – complete memory of the past
  – larger immediate memory
  – avoiding wishful thinking
  – ability to self-simulate
– greater ability than humans at organizing experience

Most required features of robot consciousness will correspond to features of human consciousness.
FEATURES OF FORMALIZED CONTEXTS

- $Ist(c, p)$, $Value(c, exp)$

- $c : p$

- $C(SherlockHolmes) : Detective(Holmes)$

- entering and leaving contexts

- introspection by transcending outermost context

- $Assuming(c, p)$
• $C(I, Now)$
What consciousness does a robot need?

- What am I doing?

\[ C(I, \text{Now}) : Driving(\text{Home,Office}) \]

- What’s my goal?

\[ C(I, \text{Now}) : Goto(\text{Office}) \]

- \( C(I, \text{Now}) : \neg \text{Know(Telephone(Mike))} \)
What Tasks Require Self-Consciousness?

Tasks NOT requiring consciousness

• Reacting directly to the environment.

• Learning direct reactions to the environment.

Tasks requiring consciousness

• Anticipating the future.

• Analyzing the past. Self-criticism.
• Speech requires introspection. Would this phrase identify this object if I were in his place?

Mechanisms of consciousness operate unconsciously.
More Tasks Requiring Consciousness

• Observe physical body.
  
  \[ ... : c(\text{Here, Now, I}) : \text{hungry} \land \text{in(\text{pen, hand})} \]

• Do I know that \textit{proposition}?
  
  \[ c(\text{Now, I}) : \neg \text{know(\text{sitting(\text{Clinton})})} \]

• Do I know what \textit{thing} is? What is it?
  
  \[ c(\text{Now, I, <pointer-to-image>}) : \text{know-what} \]

  \[ c(\text{Now, I}) : \text{is(<pointer-to-image>, jdoe}) \]
\[ c(S-Symp, I) : is(<\text{memory-image}>, j\text{doe}) \]

- Did I ever do action? When and precisely what?
- What are my goals?
- What is currently happening?
- What is the state of the actions I am currently performing?
• What are my intentions?
  \[ c(\text{Now}, I) : \text{intend}(\text{<lecture;session;lunch>}) \]

• What does my belief in \( p \) depend on?

• What are my choices for action?
  \[ c(\text{Now}, I) : \text{can(lecture)} \land \text{can(walk-out)} \]

• Can I achieve \textit{possible-goal}?

• Does my mental state up to now have property \( p \)?

• How can I plan my thinking on this problem?
Yet more Introspection

- Since I do not intend to call him again, I'll forget his telephone number—or put it in low priority storage. Packaging a proposition with a reason.

- I know how to do A and don’t know how to do B.

- Renting a cellular telephone is a new idea for me.

- I tried that, and it didn’t work. This isn’t just backtracking.

- What would I do if I were she?
Understanding

• The meaning of understanding is context dependent.

• To understand something is to have the facts and reasoning methods about it that are relevant in context.

• People who understand cars know about crankshafts.

• Fish do not understand swimming, e.g. they cannot ponder how to swim better.
• Comenici’s coach understood women’s gymnastics, but not from having done it.

• *Understanding* is an approximate concept.
Inferring Non-knowledge

Inferring non-knowledge requires special logical treatment.

• According to Gödel’s theorem, the consistency of a logical system cannot be a theorem of the system.

• Inferring that any proposition is unknown implies the system is consistent, because if the system is inconsistent, sentences are theorems.
• Gödel’s notion of relative consistency permits proofs of non-knowledge. Assume that the theory is consistent, and express this as a second order formula asserting the existence of functions and predicates with the postulated properties. To show non-knowledge of a proposition, prove that if predicates and functions exist satisfying the original theory, show that they still exist when the $\neg$ is added to the theory.

• Second order logic is the natural tool—remembering that the proof of consistency must be accomplished by the robot’s normal reasoning apparatus.
Not knowing Clinton is sitting

Theory with predicates including $sits$

$$A(P_1, \ldots, P_n, sits)$$

$$(\exists P_1', \ldots, P_n' \, sits')A(P_1', \ldots, P_n', sits')$$

expresses consistency of the theory, and

$$(\exists P_1', \ldots, P_n' \, sits')(A(P_1', \ldots, P_n', sits')$$

$$\land \neg sits'(Clinton, s))$$

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

16
Then

(8) ⊃ (9) (10)

asserts relative consistency.

\[(\exists P'_2 P'_3) A(P_1, P'_2, sits') \land \neg \text{sits}'(\text{Clinton}, s).\]

asserts it with \(P_1\) fixed. If \text{sits} doesn’t appear elsewhere, the simplest case, we get by with

\[\text{sits}' = (\lambda x \ ss)(\neg(x = \text{Clinton} \land ss = s) \lor \neg \text{sits}(x, ss))\]
Ad hoc context $c(prob)$ for a problem $prob$

- The $c(prob)$ consists mainly of a theory including facts deemed relevant to $prob$.

- $c(prob)$ is initially empty.

- $c(prob)$ is referred to from the context $c0$ in which the problem is posed by lifting relations.

- If $c(problem)$ is small enough, whether the problem is solvable in the context is definite and decideable.
• Second order logic instead of model theory keeps decisions about whether there is enough information to solve the problem within the logical language.
Relevant Work Some non-real time work is relevant to a robot examining its mental processes in real time.

- Rationalize skill—Bratko, Michie, Muggleton et. al. Shapi ro, Sternberg.

- Inductive learning systematizes and generalizes facts into predicate logic.—Muggleton
Chemistry and Logic

The interaction of chemistry and logic in humans is something we don’t need in robots. Here are some aspects of it.

- When a paranoid takes his medicine, he no longer believes the CIA is following him and influencing his thought with radio waves. When he stops taking the medicine he believes it again.

- Both the medicine and the substance to which it is antagonist are too simple to encode beliefs about the CIA.
Hormones analogous to neurotransmitters open synaptic gates to admit whole classes of beliefs into consciousness. They are analogs of similar substances and gates in animals.

It would seem that such mechanisms won't be useful for robots.
Philosophical and Psychological Remarks

The \textit{strong} design stance has advantages for philosophy:

- Gives adequacy criteria. Will the mechanism work?
- Forces a greater concreteness than is customary.
- Shows weaknesses of \textit{a priori} reasoning.
- Relative consistency evades mathematical difficulties.