WHAT CONSCIOUSNESS DOES A ROBOT NEE

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 pa
is http://www-formal.stanford.edu/consciousness.htm

APPROACHES TO ARTIFICIAL INTELLIGENCE

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

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

We aim at human level AI, and the key phenomeno the common sense informatic situation.

THE COMMON SENSE INFORMATIC SITUATIO

- Involves approximate entities.
- There is no limitation on what information may relevant. Theories must be elaboration tolerant.
- Reasoning must often be non-monotonic.

Common sense theories therefore contrast with for scientific theories and most present AI theories. Scie is embedded in common sense.

A LOGICAL ROAD TO HUMAN LEVEL AI

- Use *Drosophilas* that illustrate aspects of represen tion and reasoning problems.

- Concepts, context, circumscription, counterfactu 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 langua
  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
  some form of nonmonotonic inference, e.g. circumsc
  tion.

- Action takes place when the system infers that it sho
  do the action.

- Observation of the environment results in sentences
  memory.

- Situation calculus formalizes the relations
  $holds(p, s)$, $occurs(e, s)$ and the function $result(e, s)$ wh
  has a new situation as its value.

- Formalizing consciousness involves giving situations m
  tal components.

- Self-observation results in sentences about the syste
  state of mind.

What Introspection do Robots Need?

- *What's this?:* What ability to observe its own com
  tational state and computational processes does a ro
  need to do its tasks?

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

- *Design approach:* Asking what consciousness is nee
  gives different answers from those trying to define c
  sciouness has given.

- *Recommendation for AI*: Introspection is needed to dec
  whether to think or look, to learn from near misses
  use counterfactuals and keep pedigrees of beliefs.

- *Recommendation for psychologists and philosophers: A*
  this *direct design stance* approach to your methodolo

What is Consciousness? We consider several kinds knowledge.

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

- Unconscious knowledge can affect behavior.

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

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

- Some aspects of behavior require *decisions* of th whole system. Which way to run is an example. The decisions are made by a central mechanism.

- In logical robots, the *consciousness* is be a sub-reg
of memory containing facts and other mental entitie

- Reasoning involves the entities in consciousness a
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

Taxonomy of Consciousness

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

- Robot consciousness can be given powers people do
  have.
  - complete memory of the past
  - larger immediate memory
  - avoiding wishful thinking
  - ability to self-simulate

– greater ability than humans at organizing experien

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?

$$C(I, Now) : Driving(Home, Office)$$

- What's my goal?

$$C(I, Now) : Goto(Office)$$

- $C(I, Now) : \neg Know(Telephone(Mike))$

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 iden
  this object if I were in his place?

  Mechanisms of consciousness operate
  unconsciously.

More Tasks Requiring Consciousness

- Observe physical body.

$$\ldots : c(Here, Now, I) : hungry \wedge in(pen, hand)$$

- Do I know that *proposition*?

$$c(Now, I) : \neg know(sitting(Clinton))$$

- Do I know what *thing* is? What is it?

$$c(Now, I, <\text{pointer-to-image}>) : know\text{-}what$$

$$c(Now, I) : is(<\text{pointer-to-image}>, \text{jdoe})$$

$c(S\text{-}Symp, I) : is(<\text{memory-image}>, \text{jdoe})$

- 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 perfo ing?

- What are my intentions?

$$c(Now, I) : intend(<\text{lecture;session;lunch}>)$$

- What does my belief in p depend on?

- What are my choices for action?

$$c(Now, I) : can(lecture) \wedge can(walk\text{-}out)$$

- Can I achieve *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 t
  phone number—or put it in low priority storage. *Packa*
  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 backtra*
  ing.

- What would I do if I were she?

Understanding

- The meaning of *understanding* is *context* depende
- To understand something is to have the facts reasoning methods about it that are relevant in *context*.
- People who understand cars know about cranksha
- Fish do not understand swimming, e.g. they ca ponder how to swim better.

- Comenici's coach understood women's gymnastic
  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 log system cannot be a theorem of the system.

- Inferring that any proposition is unknown implies the s tem is consistent, because if the system is inconsisten sentences are theorems.

- Gödel's notion of *relative consistency* permits proofs
non-knowledge. Assume that the theory is consiste
and express this as a second order formula asserting
existence of functions and predicates with the postula
properties. To show non-knowledge of a propositio
prove that if predicates and functions exist satisfying
original theory, show that they still exist when the ¬
added to the theory.

- Second order logic is the natural tool—remembering t
the proof of consistency must be accomplished by
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

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

expresses the consistency of the theory with the add
assertion that Clinton is not sitting in the situation

Then

(8) $\supset$ (9)

asserts relative consistency.

$$(\exists P_2' P_3') A(P_1, P_2', sits') \wedge \neg sits'(Clinton, s).$$

asserts it with $P1$ fixed. If $sits$ doesn't appear elsewh
the simplest case, we get by with

$$sits' = (\lambda x\ ss)(\neg(x = Clinton \wedge ss = s) \vee \neg sits(x, ss))$$

Ad hoc context $c(prob)$ for a problem $prob$

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

- $c(prob)$ is initially empty.

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

- If $c(problem)$ is small enough, whether the probler
  solvable in the context is definite and decideable.

- Second order logic instead of model theory keeps
  cisions about.whether there is enough informatio
  solve the problem within the logical language.

Relevant Work Some non-real time work is relevant t
robot examining its mental processes in real time.

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

- Inductive learning systematizes and generalizes facts i
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 belie the CIA is following him and influencing his thought v radio waves. When he stops taking the medicine he lieves it again.

- Both the medicine and the substance to which it is antagonist are too simple to encode beliefs about CIA.

- Hormones analogous to neurostransmitters open syna
  gates to admit whole classes of beliefs into consciousn
  They are analogs of similar substances and gates in
  mals.

- It would seem that such mechanisms won't be usefu
  robots.

Philosophical and Psychological Remarks

The *strong* design stance has advantages for philosop

- Gives adequacy criteria. Will the mechanism work
  - Forces a greater concreteness than is customary
    - Shows weaknesses of *a priori* reasoning.
- Relative consistency evades mathematical difficulti