John McCarthy
http://www-formal.stanford.edu/jmc/
2005 November 2

THE LOGICAL ROAD TO HUMAN LEVEL AI

Will we ever reach human level AI—the main ambition of basic AI research?

Sure. Understanding intelligence is a difficult scientific problem, but lots of difficult scientific problems have been solved. There's nothing humans can do that humans can't make computers do. We, or our descendants, will have smart robot servants.

AI research should use *AI Drosophilas*, domains that are most informative about mechanisms of intelligence, not AI elephants.

Who proposed human-level AI as goal—outside of fiction?

Alan Turing was probably first—in 1947, but all the early workers in AI took human level as the goal. AI as an industrial technology with limited goals came along in the 1970s. I doubt that much of this research aimed at short term payoff is on any path to human-level AI. Indeed the researchers don't claim it.

Is there a "Moore's law" for AI? Ray Kurzweil seems to say AI performance doubles every two years.

No.

When will we get human-level AI?

Maybe 5 years. Maybe 500 years.

Will more of the same do it? The next factor of 1,000 in computer speed. More axioms in CYC of the same kind? Bigger neural nets?

No.

Most AI research today is aimed at short term payoff and avoids conceptually difficult problems.

Most likely we need fundamental new ideas. Moreover, a lot of the ideas now being pursued by hundreds of research groups are limited in scope by the remnants of behaviorist and positivist philosophy—what Steven Pinker calls *the blank slate*. I'll tell you my ideas, but most likely they are not enough.

My article *Philosophical and scientific presuppositions of logical AI*, http://www.formal.stanford.edu/jmc/phil2.html
explains what human-level AI needs in the way of philosophy.

REQUIREMENTS FOR HUMAN-LEVEL AI

An ontology adequate for stating the effects of events Examples include situations, fluents, actions and other events, and functions giving the new situations that result from events.

can be told facts e.g. the LCDs in a laptop are mounted on glass. (stated absolutely but in an implicit context).

knowledge of the common sense world—facts about dogs—3-d flexible objects, appearance including feel and smell, the effects of actions and other events.—extendable to zero-g.

the agent as one among many It knows about other agents and their likes, goals, and fears. It knows how its actions interact with those of other agents.

independence A human-level agent must not be dependent on a human to revise its concepts in face of experience, new problems, or new information. It must be at least as capable as human at reasoning about its own mental state and mental structure.

elaboration tolerance The agent must be able to take into account new information without having to be redesigned by a person.

relation between appearance and reality between 3-d objects and their 2-d projections and also with the sensation of touching them. Relation between the course of events and what we observe and do.

self-awareness The agent must regard itself as an object and as an agent and must be able to observe its own mental state.

connects reactive and deliberated action e.g. finding and removing ones keys from a pocket.

counterfactual reasoning "If another car had come over the hill when you passed, there would have been a head-on collision."

If the cop believes it, you'll be charged with reckless driving. McCarthy and Costello on "useful counterfactuals."

reasons with ill-defined entities—the purposes of the USA, the welfare of a chicken, the rocks of Mount Everest, the car that might have come over the hill.

These requirements are independent of whether the agent is logic based or an imitation of biology, e.g. a neural net.

APPROACHES TO AI

biological—imitate human, e.g. neural nets, should work eventually, but they'll have to take a more general approach.

engineering—study problems the world presents, still ahead, e.g. direct programming, genetic programming.

use logic and logical reasoning The logic approach is the most awkward—except for all the others that have been tried. Today the work with fmri makes it look like the logical and biological approaches may soon usefully interact.

WHY THE LOGIC ROAD?

If the logic road reaches human-level AI, we will have reached an understanding of how to represent the information that is available to achieve goals. A learning or evolutionary system might achieve the human-level performance without the understanding.

• Leibniz, Boole and Frege all wanted to formalize common sense. This requires methods beyond what worked to formalize mathematics—first of all formalizing nonmonotonic reasoning.

• Since 1958: McCarthy, Green, Nilsson, Fikes, Reiter, Levesque, Bacchus, Sandewall, Hayes, Lifschitz, Lin, Kowalski, Minker, Perlis, Kraus, Costello, Parmar, Amir, Morgenstern, Thielscher, Doherty, Ginsberg, McIlraith . . . —and others I have left out.

• Express facts about the world, including effects of actions and other events.

• Reason about ill-defined entities, e.g. the welfare of chickens. Thus formulas like

$Welfare(x, Result(Kill(x), s)) < Welfare(x, s)$ are sometimes needed even though $Welfare(x, s)$ is often indeterminate.

LOGIC

Describes how people think—or how people think rigorously.

The laws of deductive thought. (Boole, de Morgan, Frege, Peirce). First order logic is complete and perhaps universal.

Present mathematical logic doesn't cover all good reasoning. It does cover all guaranteed correct reasoning.

More general correct reasoning must extend logic to cover non-monotonic reasoning and probably more. Some good but non-monotonic reasoning is not guaranteed to always produce correct conclusions.

COMMON SENSE IN LOGICAL LANGUAGES—EXAMPLES

- For every boy, there's a girl who loves only him.
- $(\forall b)(\exists g)(Loves(g,b) \wedge (\exists ! b) Loves(g,b))$

This uses different sorts for boys and girls. There isn't a standard logical way of saying "loves only him".

- Block A is on Block B.

Variants: $On(A,B)$, $On(A,B,s)$, $Holds(On(A,B),s)$, $Location(A,s) = Top(B)$, $Value(Location(A),s) = Value(Top(B),s)$.

- Pat knows Mike's telephone number.

$Knows(Pat, TTelephone(MMike))$

THE COMMON SENSE INFORMATIC SITUATION

The *common sense informatic situation* is the key to human-level AI.

I have only partial information about myself and my surroundings. I don't even have a final set of concepts.

Objects of perception and thought are only partly known and are often only approximately defined.

What I think I know is subject to change and elaboration.

There is no bound on what might be relevant. The barometer *drosophila* illustrates this common sense physics. [Use a barometer to find the height of a building.]

Sometimes we (or better it) can connect a bounded informatic situation to an open informatic situation. Thus the schematic blocks world can be used to control a robot stacking real blocks.

A human-level reasoner must often do nonmonotonic reasoning.

Nevertheless, human reasoning is often very effective.

I'm in a world in which I'm a product of evolution.

THE COMMON SENSE INFORMATIC SITUATION—more

The world in which common sense operates has the following aspects.

1. Situations are snapshots of part of the world.

2. Events occur in time creating new situations. Agents' actions are events.

3. Agents have purposes they attempt to realize.

4. Processes are structures of events and situations.

5. 3-dimensional space and objects occupy regions. Embodied agents, e.g. people and physical robots are objects. Objects can move, have mass, can come apart or combine to make larger objects.

6. Knowledge of the above can only be approximate.

7. The csis includes mathematics, i.e. abstract structures and their correspondence with structures in the real world.

8. Common sense can come to include facts discovered by science. Examples are conservation of mass and conservation of volume of a liquid.

9. Scientific information and theories are imbedded in common sense information, and common sense is needed to use science.

BACKGROUND IDEAS

- epistemology (what an agent can know about the world—in general and in particular situations)

- heuristics (how to use information to achieve goals)

- declarative and procedural information

- situations

SITUATION CALCULUS

Situation calculus is a formalism dating from 1964 for representing the effects of actions and other events.

My current ideas are in *Actions and other events in situation calculus* - KR2002, available as www-formal.stanford.edu/jmc/sitcalc.html. They differ from those of Ray Reiter's 2001 book which has, however, been extended to the programming language GOLOG.

$$
\begin{array}{l}
Clear(x) \wedge Clear(l) \rightarrow At(x, l, Result(Move(x, l), s)) \\
At(y, l1) \wedge y \neq x \rightarrow At(y, l1, Result(Move(x, l), s))
\end{array}
\tag{1}
$$

Going from frame axioms to explanation closure axioms lost elaboration tolerance. The new formalism is just as concise as those based on explanation closure but, like systems using frame axioms, is *additively elaboration tolerant.*

The frame, qualification and ramification problems are identified and significantly solved in situation calculus.

There are extensions of situation calculus to concurrent and/or continuous events and actions, but the formalisms are still not entirely satisfactory.

CONCURRENCY AND PARALLELISM

- In time. *Drosophila* = Junior in Europe and Daddy in New york. When concurrent activities don't interact, the situation calculus description of the joined activities needs is the conjunction of the descriptions of the separate activities. Then the joint theory is a *conservative extension* of the separate theories. Temporal concurrency is partly done.

- In space. A situation is analyzed as composed of subpositions that are analyzed separately and then (if necessary) in interaction. *Drosophilas* are *Go* and the geometry of the Lemmings game. Spatial parallelism is hardly started.

INDIVIDUAL CONCEPTS AND PROPOSITIONS

In ordinary language concepts are objects. So be it in logic.

$$CanSpeakWith(p1, p2, Dials(p1, Telephone(p2), s))$$
$$Knows(p1, TTelephone(pp2), s) \rightarrow Cank(p1, Dial(Telephone(p2), s)$$

$$Telephone(Mike) = Telephone(Mary)$$
$$TTelephone(MMike) \neq TTelephone(MMary)$$

$$Denot(MMike) = Mike \wedge Denot(MMary) = Mary$$
$$(\forall pp)(Denot(Telephone(pp)) = Telephone(Denot(pp)))$$
$$Knows(Pat, TTelephone(MMike))$$
$$\quad \wedge \neg Knows(Pat, TTelephone(MMary))$$

CONTEXT

Relations among expressions evaluated in different contexts.

$$
\begin{aligned}
&C0 : Value(ThisLecture, I) = \text{“}JohnMcCarthy\text{''} \\
&C0 : Ist(USLegalHistory, Occupation(Holmes) = Judge) \\
&C0 : Ist(USLiteraryHistory, Occupation(Holmes) = Poet) \\
&C0 : Father(Value(USLegalHistory, Holmes)) = \\
&Value(USLiteraryHistory, Holmes)
\end{aligned}
$$

$$
\begin{aligned}
Value(C_{AFdb}, Price(GE610)) = {}& Value(C_{GEdb}, Price(GE610)) \\
&+ Value(C_{GEdb}, Price(Spares(GE610)))
\end{aligned}
$$

Can transcend outermost context, permitting introspection.

Here we use contexts as objects in a logical theory, which requires an extension to logic. The approach hasn't been popular. Too bad.

NONMONOTONIC REASONING—CIRCUMSCRIPTION

$P \leq P' \equiv (\forall x \ldots z)(P(x \ldots z) \rightarrow P'(x \ldots z))$
$P < P' \equiv P \leq P' \wedge \neg(P \equiv A')$
$Circm\{E; C; P; Z\} \equiv E(P, Z) \wedge (\forall P' \ Z')(E(P', Z') \rightarrow \neg(P' < P))$

In $Circm\{E; C; P; Z\}$, E is the axiom, C is a set of entities held constant, P is the predicate to be minimized, and Z represents predicates that can be varied in minimizing P.

$$\begin{aligned}
&\neg Ab(Aspect1(x)) \rightarrow \neg flies(x)\\
&bird(x) \rightarrow Ab(Aspect1(x))\\
&bird(x) \wedge \neg Ab(Aspect2(x)) \rightarrow flies(x)\\
&penguin(x) \rightarrow Ab(Aspect2(x))\\
&penguin(x) \wedge \neg Ab(Aspect3(x)) \rightarrow \neg flies(x)
\end{aligned}$$

Let E be the conjunction of the above sentences.

Then $Circum(E; \{bird, penguin\}; Ab; flies)$ implies

$flies(x) \equiv bird(x) \wedge \neg penguin(x)$, i.e. the things that fly are those birds that are not penguins.

frame, qualification and ramification problems

Conjecture: Simple abnormality theories aren't enough.
(No matter what the language).

Inference to a *bounded model*

SOME USES OF NONMONOTONIC REASONING

1. As a communication convention. A bird may be presumed to fly.

2. As a database convention. Flights not listed don't exist.

3. As a rule of conjecture. Only the known tools are available.

4. As a representation of a policy. The meeting is on Wednesday unless otherwise specified.

5. As a streamlined expression of probabilistic information when probabilities are near 0 or near 1. Ignore the risk of being struck by lightning.

ELABORATION TOLERANCE

Drosophila = Missionaries and Cannibals: The smallest missionary cannot be alone with the largest cannibal. One of the missionaries is Jesus Christ who can walk on water. The probability that the river is too rough is 0.1.

Additive elaboration tolerance. Just add sentences.

See www.formal.stanford.edu/jmc/elaboration.html.

Ambiguity tolerance
Drosophila = Law against conspiring to assault a federal official.

APPROXIMATE CONCEPTS AND THEORIES

Reliable logical structures on quicksand semantic foundation

Drosophila = {Mount Everest, welfare of a chicken}

No truth value to many basic propositions.
Which rocks belong to the mountain?

Definite truth value to some compound propositions whose base concepts are squishy. Did Mallory and Irvine reach the top of Everest in 1924?

HEURISTICS

Domain dependent heuristics for logical reasoning

Declarative expression of heuristics.

Wanted: General theory of special tricks

Goal: Programs that do no more search than humans do. On the 15 puzzle, Tom Costello and I got close. Shaul Markovitch got closer.

LEARNING AND DISCOVERY

Learning - what can be learned is limited by what can be represented.
Drosophila = chess

Creative solutions to problems.
Drosophila = mutilated checkerboard

Declarative information about heuristics.
Domain dependent reasoning strategies
Drosophilas = {geometry, blocks world}

Strategy in 3-d world.
Drosophila = Lemmings

Learning classifications is a very limited kind of learning problem.

Learn about reality from appearance, e.g 3-d reality from 2-d appearance. See
www-formal.stanford.edu/jmc/appearance.html for a relevant puzzle.

Learn new concepts. Stephen Muggleton's inductive logic programming is a good start.

ALL APPROACHES TO AI FACE SIMILAR PROBLEMS

Like humans AI systems must communicate in facts, not in programs or in objects. To communicate requires very little knowledge of the mental state of the recipient.

Succeeding in the common sense informatic situation requires elaboration tolerance.

It must infer reality from appearance.

Living with approximate concepts is essential

Transcending outermost context, introspection.

Nonmonotonic reasoning

INTUITIONS AND ARGUMENTS AGAINST LOGIC

• In 1975 Marvin Minsky argued that logic didn't have nonmonotonic reasoning. Nonmonotonic extensions of logic.

• The connectionist argument of 1980: Logical AI hasn't achieved human-level intelligence. Therefore, our way must be better. 25 years have elapsed, and connectionism hasn't done it either.

• Your logical language can't express X. Hence logic is inadequate. Extend the language. Getting a universal language is unsolved—requires metamathematics in the language.

• People don't reason logically, e.g. Kahneman and Twersky examples. When people reason in opposition to logic they are mistaken. Formal logic, starting with Aristotle, was first invented for communication among people and to improve human reasoning.

• Present general first order logic programs do poorly—even on problems expressed in first order logic. Better programs are needed—including metamathematical reasoning. Relying entirely on resolution was a mistake.

• Gödel showed incompleteness of first order arithmetic, and Turing showed undecideability of the halting problem. AI must work around these limitations—which also apply to human reasoning. As Turing (1930s), Gentzen (1930s) and Feferman (1962) showed, strengthening arithmetic is possible, but the theory is complicated. Some very smart people, e.g. Penrose, persistently get it wrong, perhaps because of philosophical and anti-AI biases.

QUESTIONS

What can humans do that humans can't make computers do?

What is built into newborn babies that we haven't managed to build into computer programs? Semi-permanent 3-d flexible objects.

Is there a general theory of heuristics?

First order logic is universal. Is there a general first order language? Is set theory universal enough?

What must be built in before an AI system can learn from books and by questioning people?

CAN WE MAKE A PLAN FOR HUMAN LEVEL AI?

- Study relation between appearance and reality.
www-formal.stanford.edu/jmc/appearance.html

- Extend sitcalc to full concurrency and continuous processes.

- Extend sitcalc to include strategies

- Mental sitcalc

- Reasoning within and about contexts, transcending contexts.

• Concepts as objects—as an elaboration of a theory without concepts. $Denot(TTelephone(MMike)) = Telephone(Mike)$.

• Uncertainty with and without numerical probabilities—probability of a proposition as an elaboration.

• Heavy duty axiomatic set theory. ZF with abbreviated ways of defining sets. Programs will need to invent the $E\{x\ldots\}$ used in the comprehension set former $\{x,\ldots|E\{x,\ldots\}\}$.

• Reasoning program controllable by declaratively expressed heuristics. Instead of domain dependent or reasoning style dependent logics use general logic with set theory controlled by domain dependent advice to a general reasoning program.

- All this will be difficult and needs someone young, smart, knowledgeable, and independent of the fashions in AI.
- For the rest of us: Ask oneself: Where is my work on the path to human-level AI?

AI-HARD PROBLEMS—adapted from Fanya Montalvo

Used to describe problems or subproblems in AI, to indicate that the solution presupposes a solution to the 'strong AI problem' (that is, the synthesis of a human-level intelligence). A problem that is AI-hard is, in other words, just too hard.

Examples of AI-hard problems are 'The Vision Problem' (building a system that can see as well as a human) and 'The Natural Language Problem' (building a system that can understand and speak a natural language as well as a human). These may appear to be modular, but all attempts so far (1996) to solve them have foundered on the amount of context information and 'intelligence' they seem to require.