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

THE LOGICAL ROAD TO HUMAN LEVEL

Will we ever reach human level AI—the main ambitio
AI research?

Sure. Understanding intelligence is a difficult scientifi
but lots of difficult scientific problems have been solve
nothing humans can do that humans can't make com
We, or our descendants, will have smart robot servar

AI research should use *AI Drosophilas*, domains tha
informative about mechanisms of intelligence, not AI

Who proposed human-level AI as goal—outside of fic

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

Is there a "Moore's law" for AI? Ray Kurzweil seems
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 1,00
puter speed. More axioms in CYC of the same kin
neural nets?

No.

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

Most likely we need fundamental new ideas. Moreove
the ideas now being pursued by hundreds of research
limited in scope by the remnants of behaviorist anc
philosophy—what Steven Pinker calls *the blank slat*
you my ideas, but most likely they are not enough.

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

REQUIREMENTS FOR HUMAN-LEVEL A

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

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

knowledge of the common sense world—facts abo 3-d flexible objects, appearance including feel and sm fects of actions and other events.—extendable to zer

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

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

elaboration tolerance The agent must be able to
account new information without having to be redes
person.

relation between appearance and reality between 3
and their 2-d projections and also with the sensatior
ing them. Relation between the course of events an
observe and do.

self-awareness The agent must regard itself as an
as an agent and must be able to observe its own mer

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

counterfactual reasoning "If another car had come c
when you passed, there would have been a head-on

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

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

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

APPROACHES TO AI

biological—imitate human, e.g. neural nets, should w tually, but they'll have to take a more general approa

engineering—study problems the world presents, still a direct programming, genetic programming.

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

WHY THE LOGIC ROAD?

If the logic road reaches human-level AI, we will have understanding of how to represent the information th able to achieve goals. A learning or evolutionary sys achieve the human-level performance without the unde

- Leibniz, Boole and Frege all wanted to formaliz sense. This requires methods beyond what worked to mathematics—first of all formalizing nonmonotonic r

- Since 1958: McCarthy, Green, Nilsson, Fikes, Reiter, Bacchus, Sandewall, Hayes, Lifschitz, Lin, Kowalsk

Perlis, Kraus, Costello, Parmar, Amir, Morgenstern, T
Doherty, Ginsberg, McIlraith . . . —and others I have

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

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

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

LOGIC

Describes how people think—or how people think rig

The laws of deductive thought. (Boole, de Morg
Peirce). First order logic is complete and perhaps un

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

More general correct reasoning must extend logic to
monotonic reasoning and probably more. Some goo
monotonic reasoning is not guaranteed to always produ
conclusions.

- 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
logical way of saying "loves only him".

- Block A is on Block B.

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

- Pat knows Mike's telephone number.

$Knows(Pat, TTelephone(MMike))$

THE COMMON SENSE INFORMATIC SITUAT

The *common sense informatic situation* is the key to h
AI.

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

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

What I think I know is subject to change and elabora

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

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

A human-level reasoner must often do nonmonotonic

Nevertheless, human reasoning is often very effective

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

THE COMMON SENSE INFORMATIC SITUATIO

The world in which common sense operates has the aspects.

1. Situations are snapshots of part of the world.

2. Events occur in time creating new situations. Ager 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. agents, e.g. people and physical robots are object can move, have mass, can come apart or combin larger objects.

6. Knowledge of the above can only be approximate

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

8. Common sense can come to include facts discove ence. Examples are conservation of mass and co of volume of a liquid.

9. Scientific information and theories are imbedded i sense information, and common sense is needed ence.

BACKGROUND IDEAS

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

- heuristics (how to use information to achieve goa

- declarative and procedural information

- situations

SITUATION CALCULUS

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

My current ideas are in *Actions and other events in sit*
culus - KR2002, available as www-formal.stanford.edu
They differ from those of Ray Reiter's 2001 book
however, been extended to the programming language

$$Clear(x) \wedge Clear(l) \rightarrow At(x, l, Result(Move(x, l),$$
$$At(y, l1) \wedge y \neq x \rightarrow At(y, l1, Result(Move(x, l), s$$

Going from frame axioms to explanation closure axiom
oration tolerance. The new formalism is just as concis
based on explanation closure but, like systems using
ioms, is *additively elaboration tolerant.*

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

There are extensions of situation calculus to concurre
continuous events and actions, but the formalisms a
entirely satisfactory.

CONCURRENCY AND PARALLELISM

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

- In space. A situation is analyzed as composed c tions that are analyzed separately and then (if ne interaction. *Drosophilas* are *Go* and the geome Lemmings game. Spatial parallelism is hardly star

INDIVIDUAL CONCEPTS AND PROPOSITIO

In ordinary language concepts are objects. So be it in

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

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

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

CONTEXT

Relations among expressions evaluated in different co

$C0 : Value(ThisLecture, I) = \text{“}JohnMcCarthy''$
$C0 : Ist(USLegalHistory, Occupation(Holmes) = J$
$C0 : Ist(USLiteraryHistory, Occupation(Holmes) =$
$C0 : Father(Value(USLegalHistory, Holmes)) =$
$Value(USLiteraryHistory, Holmes)$

$$Value(C_{AFdb}, Price(GE610)) = Value(C_{GEdb}, Price($$
$$+Value(C_{GEdb}, Price(Spares(GE610)))$$

Can transcend outermost context, permitting introsp

Here we use contexts as objects in a logical theory, whi
an extension to logic. The approach hasn't been pop
bad.

NONMONOTONIC REASONING—CIRCUMSCRI

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

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

$$\neg Ab(Aspect1(x)) \rightarrow \neg flies(x)$$
$$bird(x) \rightarrow Ab(Aspect1(x))$$
$$bird(x) \wedge \neg Ab(Aspect2(x)) \rightarrow flies(x)$$
$$penguin(x) \rightarrow Ab(Aspect2(x))$$
$$penguin(x) \wedge \neg Ab(Aspect3(x)) \rightarrow \neg flies(x)$$

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 fl
birds that are not penguins.

frame, qualification and ramification problems

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

Inference to a *bounded model*

SOME USES OF NONMONOTONIC REASON

1. As a communication convention. A bird may be pr fly.

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

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

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

5. As a streamlined expression of probabilistic informa probabilities are near 0 or near 1. Ignore the risk of be by lightning.

ELABORATION TOLERANCE

Drosophila = Missionaries and Cannibals: The smalle
ary cannot be alone with the largest cannibal. One o
sionaries is Jesus Christ who can walk on water. The
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 fede

APPROXIMATE CONCEPTS AND THEORI

Reliable logical structures on quicksand semantic fou

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 w
concepts are squishy. Did Mallory and Irvine reach
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 humar
the 15 puzzle, Tom Costello and I got close. Shaul I
got closer.

LEARNING AND DISCOVERY

Learning - what can be learned is limited by what car
sented.
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 learnin

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

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

ALL APPROACHES TO AI FACE SIMILAR PRO

Like humans AI systems must communicate in facts,
grams or in objects. To communicate requires very li
edge of the mental state of the recipient.

Succeeding in the common sense informatic situatio
elaboration tolerance.

It must infer reality from appearance.

Living with approximate concepts is essential

Transcending outermost context, introspection.

Nonmonotonic reasoning

INTUITIONS AND ARGUMENTS AGAINST LO

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

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

• Your logical language can't express X. Hence logi quate. Extend the language. Getting a universal la unsolved—requires metamathematics in the language

• People don't reason logically, e.g. Kahneman an
examples. When people reason in opposition to logi
mistaken. Formal logic, starting with Aristotle, wa
vented for communication among people and to impr
reasoning.

• Present general first order logic programs do po
on problems expressed in first order logic. Better
are needed—including metamathematical reasoning. F
tirely on resolution was a mistake.

• Gödel showed incompleteness of first order arithmeti
ing showed undecideability of the halting problem. AI

around these limitations—which also apply to huma
ing. As Turing (1930s), Gentzen (1930s) and Feferm
showed, strengthening arithmetic is possible, but th
complicated. Some very smart people, e.g. Penrose, p
get it wrong, perhaps because of philosophical and ant

QUESTIONS

What can humans do that humans can't make comp

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

Is there a general theory of heuristics?

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

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

CAN WE MAKE A PLAN FOR HUMAN LEVE

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

- Extend sitcalc to full concurrency and continuous p

- Extend sitcalc to include strategies

- Mental sitcalc

- Reasoning within and about contexts, transcending

• Concepts as objects—as an elaboration of a theo
concepts. $Denot(TTelephone(MMike)) = Telephone(M$

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

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

• Reasoning program controllable by declaratively expre
tics. Instead of domain dependent or reasoning style

logics use general logic with set theory controlled
dependent advice to a general reasoning program.

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

AI-HARD PROBLEMS—adapted from Fanya Mo

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

Examples of AI-hard problems are 'The Vision Proble
ing a system that can see as well as a human) and 'T
Language Problem' (building a system that can unde
speak a natural language as well as a human). Thes
pear to be modular, but all attempts so far (1996) to
have foundered on the amount of context informatio
telligence' they seem to require.