THE LOGICAL ROAD TO HUMAN LEVEL

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

Sure. Understanding intelligence is a difficult scientific problem, but lots of difficult scientific problems have been solved. 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 informative about mechanisms of intelligence, not AI.
Who proposed human-level AI as goal—outside of fiction?

Alan Turing was probably first—in 1947, but all the early work in AI took human level as the goal. AI as an industrial technology with limited goals came along in the 1970s. I doubt a significant fraction of this research aimed at short term payoff is on anything approaching human-level AI. Indeed the researchers don’t claim it.

Is there a “Moore’s law” for AI? Ray Kurzweil seems to think so, but 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 computer speed. More axioms in CYC of the same kind as neural nets?

No.

Most AI research today is aimed at short term payoff and conceptually difficult problems.
Most likely we need fundamental new ideas. Moreover, the ideas now being pursued by hundreds of research groups are limited in scope by the remnants of behaviorist and cognitive psychology—what Steven Pinker calls the blank slate. I'll try to explain to you my ideas, but most likely they are not enough.

REQUIREMENTS FOR HUMAN-LEVEL AI

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

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

knowledge of the common sense world—facts about 3-d flexible objects, appearance including feel and smell, 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 compare 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 data, or new information. It must be at least as capable as a human in reasoning about its own mental state and mental structure.

elaboration tolerance The agent must be able to account for new information without having to be redesigned or person.
relation between appearance and reality between 3-d and their 2-d projections and also with the sensations arising them. Relation between the course of events and what we observe and do.

self-awareness The agent must regard itself as an object as an agent and must be able to observe its own mental connectivity reactive and deliberated action e.g. finding removing ones keys from a pocket.

counterfactual reasoning “If another car had come over when you passed, there would have been a head-on
If the cop believes it, you’ll be charged with reckless behavior. As McCarthy and Costello on “useful counterfactuals.”

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

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

use logic and logical reasoning The logic approach is awkward—except for all the others that have been tried; the work with fMRI makes it look like the logical and other 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 enables us to achieve goals. A learning or evolutionary system should be able to achieve the human-level performance without the understanding.

- Leibniz, Boole and Frege all wanted to formalize human understanding. This requires methods beyond what worked to formalize mathematics—first of all formalizing nonmonotonic reasoning.

- Since 1958: McCarthy, Green, Nilsson, Fikes, Reiter, Bacchus, Sandewall, Hayes, Lifschitz, Lin, Kowalski
• Express facts about the world, including effects of actions and other events.

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

\[ \text{Welfare}(x, \text{Result}(\text{Kill}(x), s)) < \text{Welfare}(x, s) \]

are sometimes meaningful even though \( \text{Welfare}(x, s) \) is often indeterminate.
LOGIC

Describes how people think—or how people think right?

The laws of deductive thought. (Boole, de Morgan, 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 monotonic reasoning and probably more. Some good monotonic reasoning is not guaranteed to always produce conclusions.
• For every boy, there’s a girl who loves only him.

$(\forall b)(\exists g)(Loves(g, b) \land (\exists! b)Loves(g, b))$

This uses different sorts for boys and girls. There isn’t a logical way of saying “loves only him”.

• Block A is on Block B.

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

• Pat knows Mike’s telephone number.

$Knows(Pat, Telephone(Mike))$
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 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 *drosophila* illustrates this common sense physics. [Use a meter 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 real

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 SITUATION

The world in which common sense operates has these aspects.

1. Situations are snapshots of part of the world.

2. Events occur in time creating new situations. Agents 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 objects. They can move, have mass, can come apart or combine to form 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 through experience. 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 understand it.
BACKGROUND IDEAS

- epistemology (what an agent can know about the 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
modeling the effects of actions and other events.

My current ideas are in Actions and other events in sit-
They differ from those of Ray Reiter’s 2001 book which
however, been extended to the programming language

\[
\begin{align*}
\text{Clear}(x) \land \text{Clear}(l) & \rightarrow \text{At}(x, l, \text{Result}(\text{Move}(x, l), s)) \\
\text{At}(y, l_1) \land y \neq x & \rightarrow \text{At}(y, l_1, \text{Result}(\text{Move}(x, l), s))
\end{align*}
\]
Going from frame axioms to explanation closure axioms is elaboration tolerance. The new formalism is just as concise based on explanation closure but, like systems using axioms, is *additively elaboration tolerant*.

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

There are extensions of situation calculus to concurrent continuous events and actions, but the formalisms are entirely satisfactory.
CONCURRENCY AND PARALLELISM

• In time. *Drosophila* = Junior in Europe and Daddy in New York. When concurrent activities don’t interact, the calculus description of the joined activities needs to be the conjunction of the descriptions of the separate activities, and the joint theory is a *conservative extension* of the component theories. Temporal concurrency is partly done.

• In space. A situation is analyzed as composed of situations that are analyzed separately and then (if necessary) interaction. *Drosophilas* are *Go* and the geometry of Lemmings game. Spatial parallelism is hardly started.
In ordinary language concepts are objects. So be it in

\[\text{CanSpeakWith}(p_1, p_2, \text{Dials}(p_1, \text{Telephone}(p_2), s))\]
\[\text{Knows}(p_1, \text{TTelephone}(p_2), s) \rightarrow \text{Cank}(p_1, \text{Dial}(\text{Telephone}(p_2)))\]

\[\text{Telephone}(\text{Mike}) = \text{Telephone}(\text{Mary})\]
\[\text{TTelephone}(\text{MMike}) \neq \text{TTelephone}(\text{MMary})\]

\[\text{Denot}(\text{MMike}) = \text{Mike} \land \text{Denot}(\text{MMary}) = \text{Mary}\]
\[(\forall pp)(\text{Denot}(\text{Telephone}(pp)) = \text{Telephone}(\text{Denot}(pp)))\]
\[\text{Knows}(\text{Pat}, \text{TTelephone}(\text{MMike}))\]
\[\land \neg \text{Knows}(\text{Pat}, \text{TTelephone}(\text{MMary}))\]
Relations among expressions evaluated in different contexts.

C0 : Value(ThisLecture, I) = "John McCarthy"
C0 : Ist(USLegalHistory, Occupation(Holmes)) = Judge
C0 : Ist(USLiteraryHistory, Occupation(Holmes)) = Judge
C0 : Father(Value(USLegalHistory, Holmes)) = Value(USLiteraryHistory, Holmes)

\[ Value(C_{AFdb}, Price(GE610)) = Value(C_{GEdb}, Price(GE610) + Value(C_{GEdb}, Price(Spares(GE610)))) \]

Can transcend outermost context, permitting introspection.

Here we use contexts as objects in a logical theory, which is an extension to logic. The approach hasn’t been popular, but it isn’t bad.
NONMONOTONIC REASONING—CIRCUMSCRIPTION

\[ P \leq P' \equiv (\forall x \ldots z)(P(x \ldots z) \rightarrow P'(x \ldots z)) \]
\[ P < P' \equiv P \leq P' \land \neg(P \equiv A') \]
\[ \text{Circm}\{E; C; P; Z\} \equiv E(P, Z) \land (\forall P' Z')(E(P', Z') \rightarrow \neg\text{In}\text{Circm}\{E; C; P; Z\}) \]

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

\[ \neg\text{Ab}(\text{Aspect1}(x)) \rightarrow \neg\text{flies}(x) \]
\[ \text{bird}(x) \rightarrow \text{Ab}(\text{Aspect1}(x)) \]
\[ \text{bird}(x) \land \neg\text{Ab}(\text{Aspect2}(x)) \rightarrow \text{flies}(x) \]
\[ \text{penguin}(x) \rightarrow \text{Ab}(\text{Aspect2}(x)) \]
\[ \text{penguin}(x) \land \neg\text{Ab}(\text{Aspect3}(x)) \rightarrow \neg\text{flies}(x) \]
Let $E$ be the conjunction of the above sentences.

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

$$flies(x) \equiv bird(x) \land \neg penguin(x),$$

i.e. the things that fly are 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 \textit{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 used.

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

5. As a streamlined expression of probabilistic information. If probabilities are near 0 or near 1, ignore the risk of being struck by lightning.
ELABORATION TOLERANCE

*Drosophila* = Missionaries and Cannibals: The smaller 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
Reliable logical structures on quicksand semantic foundation

\[ Drosophila = \{ \text{Mount Everest, welfare of a chicken} \} \]

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

Definite truth value to some compound propositions whose 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 humans. On the 15 puzzle, Tom Costello and I got close. Shaul Ma got closer.
LEARNING AND DISCOVERY

Learning - what can be learned is limited by what can be presented.
\(Drosophila = \) chess

Creative solutions to problems.
\(Drosophila = \) mutilated checkerboard

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

Strategy in 3-d world.
\(Drosophila = \) Lemmings
Learning classifications is a very limited kind of learning.

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

Learn new concepts. Stephen Muggleton’s inductive programming is a good start.
ALL APPROACHES TO AI FACE SIMILAR PROBLEMS

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

Succeeding in the common sense informatic situation 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 flawed. Years have elapsed, and connectionism hasn’t done it.

- Your logical language can’t express X. Hence logic is inadequate. Extend the language. Getting a universal language—unsolved—requires metamathematics in the language.
• People don’t reason logically, e.g. Kahneman and examples. When people reason in opposition to logic, it is mistaken. Formal logic, starting with Aristotle, was invented for communication among people and to improve reasoning.

• Present general first order logic programs do poorly 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, showing showed undecideability of the halting problem. AI
around these limitations—which also apply to human thinking. As Turing (1930s), Gentzen (1930s) and Feferman showed, strengthening arithmetic is possible, but the process is complicated. Some very smart people, e.g. Penrose, perhaps get it wrong, perhaps because of philosophical and anti
QUESTIONS

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

What is built into newborn babies that we haven’t yet built into computer programs? Semi-permanent 3D 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 people and by questioning people?
CAN WE MAKE A PLAN FOR HUMAN LEVEL

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

- Extend sitcalc to full concurrency and continuous problems

- Extend sitcalc to include strategies

- Mental sitcalc

- Reasoning within and about contexts, transcending
• Concepts as objects—as an elaboration of a theory of concepts. \( \text{Denot}(TT_{\text{elephone}}(M_{\text{Mike}})) = T_{\text{elephone}}(M) \)

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

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

• Reasoning program controllable by declaratively expressivestatics. 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, smudgeable, and independent of the fashions in AI.
- For the rest of us: Ask oneself: Where is my work on getting to human-level AI?
AI-HARD PROBLEMS—adapted from Fanya Montalvo

Used to describe problems or subproblems in AI, to indicate the solution presupposes a solution to the ‘strong AI’ (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 Problem’ (building a system that can see as well as a human) and ‘The Language Problem’ (building a system that can understand and speak a natural language as well as a human). These appear to be modular, but all attempts so far (1996) to solve them have foundered on the amount of context information ‘intelligence’ they seem to require.