FROM HERE TO HUMAN-LEVEL AI

John McCarthy
Computer Science Department
Stanford University
Stanford, CA 94305
jmc@cs.stanford.edu
http://www-formal.stanford.edu/jmc/

Abstract

It is not surprising that reaching human-level AI has proved to be difficult and progress has been slow—though there has been important progress. The slowness and the demand to exploit what has been discovered has led many to mistakenly redefine AI, sometimes in ways that preclude human-level AI—by relegating to humans parts of the task that human-level computer programs would have to do. In the terminology of this paper, it amounts to settling for a *bounded informatic situation* instead of the more general *common sense informatic situation*.

Overcoming the "brittleness" of present AI systems and reaching human-level AI requires programs that deal with the *common sense informatic situation*—in which the phenomena to be taken into account in achieving a goal are not fixed in advance.

We discuss reaching human-level AI, emphasizing logical AI and especially emphasizing representation problems of information and of reasoning. Ideas for reasoning in the common sense informatic situation include nonmonotonic reasoning, approximate concepts, formalized contexts and introspection.

1 What is Human-Level AI?

The first scientific discussion of human level machine intelligence was apparently by Alan Turing in the lecture [Turing, 1947]. The notion was amplified as a goal in [Turing, 1950], but at least the latter paper did not say what would have to be done to achieve the goal.

Allen Newell and Herbert Simon in 1954 were the first people to make a start on programming computers for general intelligence. They were over-optimistic, because their idea of what has to be done to achieve human-level intelligence was inadequate. The *General Problem Solver* (GPS) took general problem solving to be the task of transforming one expression into another using an allowed set of transformations.

Many tasks that humans can do, humans cannot yet make computers do. There are two approaches to human-level AI, but each presents difficulties. It isn't a question of deciding between them, because each should eventually succeed; it is more a race.

1. If we understood enough about how the human intellect works, we could simulate

it. However, we don't have have sufficient ability to observe ourselves or others to understand directly how our intellects work. Understanding the human brain well enough to imitate its function therefore requires theoretical and experimental success in psychology and neurophysiology. [1] See [Newell and Simon, 1972] for the beginning of the information processing approach to psychology.

2. To the extent that we understand the problems achieving goals in the world presents to intelligence we can write intelligent programs. That's what this article is about.

What problems does the world present to intelligence? More narrowly, we consider the problems it would present to a human scale robot faced with the problems humans might be inclined to relegate to sufficiently intelligent robots. The physical world of a robot contains middle sized objects about which its sensory apparatus can obtain only partial information quite inadequate to fully determne the effects of its future actions. Its mental world includes its interactions with people and also meta-information about the information it has or can obtain.

Our approach is based on what we call the *common sense informatic situation.* In order to explain the common sense informatic situation, we contrast it with the *bounded informatic situation* that characterizes both formal scientific theories and almost all (maybe all) experimental work in AI done so far.[2]

[1]Recent work with positron emission tomography has identified areas of the brain that consume more glucose when a person is doing mental arithmetic. This knowledge will help build AI systems only when it becomes possible to observe what is going on in these areas during mental arithmetic.

[2]The textbook [David Poole and Goebel, 1998] puts it this way. "To get human-level computational intelligence it must be the agent itself that decides how to divide up the world, and which relationships to reason about.

A formal theory in the physical sciences deals with a *bounded informatic situation.* Scientists decide informally in advance what phenomena to take into account. For example, much celestial mechanics is done within the Newtonian gravitational theory and does not take into account possible additional effects such as outgassing from a comet or electromagnetic forces exerted by the solar wind. If more phenomena are to be considered, a person must make a new theory. Probabilistic and fuzzy uncertainties can still fit into a bounded informatic system; it is only necessary that the set of possibilities (sample space) be bounded.

Most AI formalisms also work only in a bounded informatic situation. What phenomena to take into account is decided by a person before the formal theory is constructed. With such restrictions, much of the reasoning can be monotonic, but such systems cannot reach human level ability. For that, the machine will have to decide for itself what information is relevant. When a *bounded informatic system* is appropriate, the system must construct or choose a limited *context* containing a suitable theory whose predicates and functions connect to the machine's inputs and outputs in an appropriate way. The logical tool for this is *nonmonotonic* reasoning.

2 The Common Sense Informatic Situation

Contention: The key to reaching human-level AI is making systems that operate successfully in the common sense informatic situation.

In general a thinking human is in what we call the *common sense informatic situation* first discussed in[3] [McCarthy, 1989]. It is more general than any *bounded informatic situation.* The known facts are incomplete, and there is no *a priori* limitation on what facts are rel-

[3]http://www-formal.stanford.edu/jmc/ailogic.html

evant. It may not even be decided in advance what phenomena are to be taken into account. The consequences of actions cannot be fully determined. The *common sense informatic situation* necessitates the use of *approximate concepts* that cannot be fully defined and the use of *approximate theories* involving them. It also requires *nonmonotonic* reasoning in reaching conclusions.

The common sense informatic situation also includes some knowledge about the system's mental state.

A nice example of the common sense informatic situation is illustrated by an article in the *American Journal of Physics* some years ago. It discussed grading answers to a physics problem. The exam problem is to find the height of a building using a barometer. The intended solution is to measure the air pressure at the top and bottom of the building and multiply the difference by the ratio of the density of mercury to the density of air.

However, other answers may be offered. (1) drop the barometer from the top of the building and measure the time before it hits the ground. (2) Measure the height and length of the shadow of the barometer and measure the length of the shadow of the building. (3) Rappel down the building using the barometer as a measuring rod. (4) Lower the barometer on a string till it reaches the ground and measure the string. (5) Offer the barometer to the janitor of the building in exchange for information about the height. (6) Ignore the barometer, count the stories of the building and multiply by ten feet.

Clearly it is not possible to bound in advance the common sense knowledge of the world that may be relevant to grading the problem. Grading some of the solutions requires knowledge of the formalisms of physics and the physical facts about the earth, e.g. the law of falling bodies or the variation of air pressure with altitude. However, in every case, the physics knowledge is embedded in common sense knowledge. Thus before one can use Galileo's law of falling bodies $s = \frac{1}{2}gt^2$, one needs common sense information about buildings, their shapes and their roofs.

Bounded informatic situations are obtained by nonmonotonically inferring that only the phenomena that somehow appear to be relevant are relevant. In the barometer example, the student was expected to infer that the barometer was only to be used in the conventional way for measuring air pressure. For example, a reasoning system might do this by applying circumscription to a predicate *relevant* in a formalism containing also metalinguistic information, e.g. that this was a problem assigned in a physics course. Formalizing relevance in a useful way promises to be difficult.

Common sense facts and common sense reasoning are necessarily imprecise. The imprecision necessitated by the common sense informatic situation applies to computer programs as well as to people.

Some kinds of imprecision can be represented numerically and have been explored with the aid of Bayesian networks, fuzzy logic and similar formalisms. This is in addition to the study of approximation in numerical analysis and the physical sciences.

3 The Use of Mathematical Logic

What about mathematical logical languages?

Mathematical logic was devised to formalize precise facts and correct reasoning. Its founders, Leibniz, Boole and Frege, hoped to use it for common sense facts and reasoning, not realizing that the imprecision of concepts used in common sense language was often a necessary feature and not always a bug. The biggest success of mathematical logic was in formalizing mathematical theories. Since the common sense informatic situation requires using imprecise facts and imprecise reasoning, the use of mathematical logic for common sense has had limited success. This has caused many people to give up. Gradually, extended logical languages and even extended forms of mathematical logic are being invented and developed.

It is necessary to distinguish between mathematical logic and particular mathematical logical languages. Particular logical languages are determined by a particular choice of concepts and the predicate and function symbols to represent them. Failure to make the distinction has often led to error. When a particular logical language has been shown inadequate for some purpose, some people have concluded that logic is inadequate. Different concepts and different predicate and function symbols might still succeed. In the words of the drive-in movie critic of Grapevine, Texas, "I'm surprised I have to explain this stuff."

The pessimists about logic or some particular set of predicates might try to prove a theorem about its inadequacies for expressing common sense.[4]

Since it seems clear that humans don't use logic as a basic internal representation formalism, maybe something else will work better for AI. Researchers have been trying to find this something else since the 1950s but still haven't succeeded in getting anything that is ready to be applied to the common sense informatic situation. Maybe they will eventually succeed. However, I think the problems listed in the later sections of this article will apply to any approach to human-level AI.

Mathematical logic has been concerned with how people ought to think rather than how people do think. We who use logic as a basic AI formalism make programs reason logically. However, we have to extend logic and extend the programs that use it in various ways.

One important extension was the development of modal logic starting in the 1920s and using it to treat modalities like knowledge, belief and obligation. Modalities can be treated either by using modal logic or by reifying concepts and sentences within the standard logic. My opinion is that reification in standard logic is more powerful and will work better.

A second extension was the formalization of nonmonotonic reasoning beginning in the late 1970s—with circumscription and default logic and their variants as the major proposals. Nonmonotonic logic has been studied both as pure mathematics and in application to AI problems, most prominently to the formalization of action and causality. Several variants of the major formalisms have been devised.

Success so far has been moderate, and it isn't clear whether greater success can be obtained by changing the the concepts and their representation by predicate and function symbols or by varying the nonmonotonic formalism. [5]

We need to distinguish the actual use of logic from what Allen Newell, [Newell, 1981] and [Newell, 1993], calls the logic level and which was also proposed in [McCarthy, 1979][6].

4 Approximate Concepts and Approximate Theories

Other kinds of imprecision are more fundamental for intelligence than numerical imprecision. Many phenomena in the world are appropriately described in terms of *approximate concepts*. Although the concepts are imprecise, many statements using them have precise truth values. We offer two examples: the con-

[4]Gödel's theorem is not relevant to this, because the question is not one of decideability or of characterizing truth.

[5]One referee for KR96 foolishly and arrogantly proposed rejecting a paper on the grounds that the inadequacy of circumscription for representing action was known.

[6]http://www-formal.stanford.edu/jmc/ascribing.html

cept of Mount Everest and the concept of the welfare of a chicken. The exact pieces of rock and ice that constitute Mount Everest are unclear. For many rocks, there is no *truth of the matter* as to whether it is part of Mount Everest. Nevertheless, it is true without qualification that Edmund Hillary and Tenzing Norgay climbed Mount Everest in 1953 and that John McCarthy never set foot on it.

The point of this example is that it is possible and even common to have a solid knowledge structure from which solid conclusions can be inferred based on a foundation built on the quicksand of approximate concepts without definite extensions.

As for the chicken, it is clear that feeding it helps it and wringing its neck harms it, but it is unclear what its welfare consists of over the course of the decade from the time of its hatching. Is it better off leading a life of poultry luxury and eventually being slaughtered or would it be better off escaping the chicken yard and taking its chances on starvation and foxes? There is no *truth of the matter* to be determined by careful investigation of chickens. **When a concept is inherently approximate, it is a waste of time to try to give it a precise definition.** Indeed different efforts to define such a concept precisely will lead to different results—if any.

Most human common sense knowledge involves approximate concepts, and reaching human-level AI requires a satisfactory way of representing information involving approximate concepts.

5 Nonmonotonic Reasoning

Common sense reasoning is also imprecise in that it draws conclusions that might not be made if there were more information. Thus common sense reasoning is *nonmonotonic*. I will not go into the details of any of the proposals for handling nonmonotonic reasoning.

In particular, getting from the common sense informatic situation to a bounded informatic situation needs nonmonotonic reasoning.

6 Elaboration Tolerance

Human abilities in the common sense informatic situation also include what may be called *elaboration tolerance*—the ability to elaborate a statement of some facts without having to start all over. Thus when we begin to think about a problem, e.g. determining the height of a building, we form a bounded context and try to solve the problem within it. However, at any time more facts can be added, e.g. about the precision with which the time for the barometer to fall can be estimated using a stop watch and also the possibilities of acquiring a stop watch.

Elaboration Tolerance[7] discusses about 25 elaborations of the Missionaries and Cannibals problem.

What I have so far said so far about approximate concepts, nonmonotonic reasoning and elaboration tolerance is independent of whether mathematical logic, human language or some other formalism is used.

In my opinion, the best AI results so far have been obtained using and extending mathematical logic.

7 Formalization of Context

A third extension of mathematical logic involves **formalizing the notion of context**[8] [McCarthy, 1993]. Notice that when logical theories are used in human communication and study, the theory is used in a context which people can discuss from the outside. If computers are to have this facility and are to

[7] http://www-formal.stanford.edu/jmc/elaboration.html

[8] http://www-formal.stanford.edu/jmc/context.html

work within logic, then the "outer" logical language needs names for contexts and sentences giving their relations and a way of entering a context. Clearly human-level AI requires reasoning about context.

Human-level AI also requires the ability to *transcend* the outermost context the system has used so far. Besides in [McCarthy, 1993], this is also discussed in **Making Robots Conscious of their Mental States**[9] [McCarthy, 1996].

Further work includes [Buvač, 1996] and [Buvač et al., 1995].

8 Reasoning about Events—Especially Actions

Reasoning about actions has been a major AI activity, but this paper will not discuss my or other people's current approaches, concentrating instead on the long range problem of reaching human level capability. We regard actions as particular kinds of events and therefore propose subsuming reasoning about actions under the heading of reasoning about events.

Most reasoning about events has concerned determining the effects of an explicitly given sequence of actions by a single actor. Within this framework various problems have been studied.

- The frame problem concerns not having to state what does not change when an event occurs.

- The qualification problem concerns not having to state all the preconditions of an action or other event. The point is both to limit the set of preconditions and also to jump to the conclusion that unstated others will be fulfilled unless there is evidence to the contrary. For example, wearing clothes is a precondition for airline travel, but the travel agent will not tell his customer to be sure and wear clothes.

- The ramification problem concerns how to treat side-effects of events other than the principal effect mentioned in the event description.

Each of these involves elaboration tolerance, e.g. adding descriptions of the effects of additional events without having to change the descriptions of the events already described. When I wrote about **applications of circumscription to formalizing common sense**[10] [McCarthy, 1986], I hoped that a *simple abnormality theory* would suffice for all of them. That didn't work out when I tried it, but I still think a common nonmonotonic reasoning mechanism will work. Tom Costello's draft *"The Expressive Power of Circumscripttion"*[11] argues that simple abnormality theories have the same expressive power as more elaborate nonmonotonic formalisms that have been proposed.

Human level intelligence requires reasoning about strategies of action, i.e. action programs. It also requires considering multiple actors and also concurrent events and continuous events. Clearly we have a long way to go.

Some of these points are discussed in a draft on narrative[12] [McCarthy, 1995].

9 Introspection

People have a limited ability to observe their own mental processes. For many intellectual tasks introspection is irrelevant. However, it is at least relevant for evaluating how one is using one's own thinking time. Human-level AI will require introspective ability.

[9] http://www-formal.stanford.edu/jmc/consciousness.html

[10] http://www-formal.stanford.edu/jmc/applications.html

[11] http://www-formal.stanford.edu/tjc/expressive.html

[12] http://www-formal.stanford.edu/jmc/narrative.html

That robots also need introspection[13] is argued and how to do it is discussed in [McCarthy, 1996].

10 Heuristics

The largest qualitative gap between human performance and computer performance is in the area of heuristics, even though the gap is disguised in many applications by the millions-fold speed advantage of computers. The general purpose theorem proving programs run very slowly, and the special purpose programs are very specialized in their heuristics.

I think the problem lies in our present inability to give programs domain and problem dependent heuristic advice. In my Advice Taker paper[14] [McCarthy, 1959] I advertised that the Advice Taker would express its heuristics declaratively. Maybe that will work, but neither I nor anyone else has been able to get a start on the problem in the ensuing almost 40 years. Josefina Sierra-Santibanez reports on some progress in a forthcoming article.

Another possibility is to express the advice in a *procedure modification language*, i.e. to extend elaboration tolerance to programs. Of course, every kind of modularity, e.g. object orientation, gives some elaboration tolerance, but these devices haven't been good enough.

Ideally, a general purpose reasoning system would be able to accept advice permitting it to run at a fixed ratio speed of speeds to a special purpose program, e.g. at 1/20 th the speed.

11 Summary

Conclusion: Between us and human-level intelligence lie many problems. They can be summarized as that of succeeding in the *common sense informatic situation.*

The problems include:

common sense knowledge of the world Many important aspects of what this knowledge is in and how it can be represented are still unsolved questions. This is particularly true of knowledge of the effects of actions and other events.

epistemologically adequate languages These are languages for expressing **what a person or robot can actually learn about the world**[15] [McCarthy and Hayes, 1969].

elaboration tolerance What a person knows can be elaborated without starting all over.

nonmonotonic reasoning Perhaps new systems are needed.

contexts as objects This subject is just beginning. See the references of section 7.

introspection AI systems will need to examine their own internal states.

action The present puzzles of formalizing action should admit a uniform solution.

I doubt that a human-level intelligent program will have structures corresponding to all these entities and to the others that might have been listed. A generally intelligent logical program probably needs only its monotonic and nonmonotonic reasoning mechanisms plus mechanisms for entering and leaving contexts. The rest are handled by particular functions and predicates.

[13] http://www-formal.stanford.edu/jmc/consciousness.html
[14] http://www-formal.stanford.edu/jmc/mcc59.html
[15] http://www-formal.stanford.edu/jmc/mcchay69.html

12 Remarks and Acknowledgements

1. To what extent will all these problems have to be faced explicitly by people working with neural nets and connectionist systems? The systems I know about are too primitive for the problems even to arise. However, more ambitious systems will inhabit the common sense informatic situation. They will have to be elaboration tolerant and will require some kind of mental model of the consequences of actions.

2.

3. I got useful suggestions from Eyal Amir, Saša Buvač and Tom Costello.

4. Some additional relevant papers are in my book [McCarthy, 1990] and on my Web site[16].

5. My understanding that I should prepare a printable version of this invited talk came rather late. I expect that both the spoken version and the 1996 November Web version will have better explanations of the important concepts.

6. This work was partly supported by ARPA (ONR) grant N00014-94-1-0775.

13 Conclusion

Many will find dismayingly large the list of tasks that must be accomplished in order to to reach human-level logical intelligence. Perhaps fewer but more powerful ideas would simplify the list. Others will claim that a system that evolves intelligence as life does will be more straightforward to build. Maybe, but the advocates of that approach have been at it as long as we have and still aren't even close.

So it's a race.

It will be much more scientifically satisfying to understand human level artificial intelligence logically than just achieve it by a computerized evolutionary process that produced an intelligent but incomprehensible result. In fact, the logical approach would be worth pursuing even if the intellectually lazy evolutionary approach won the race.

References

[Buvač, 1996] Buvač, S. (1996). Quantificational logic of context. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*.

[Buvač et al., 1995] Buvač, S., Buvač, V., and Mason, I. A. (1995). Metamathematics of contexts. *Fundamenta Informaticae*, 23(3).

[David Poole and Goebel, 1998] David Poole, A. M. and Goebel, R. (1998). *Computational Intelligence*. Oxford.

[McCarthy, 1959] McCarthy, J. (1959). Programs with Common Sense[17]. In *Mechanisation of Thought Processes, Proceedings of the Symposium of the National Physics Laboratory*, pages 77–84, London, U.K. Her Majesty's Stationery Office. Reprinted in McC90.

[McCarthy, 1979] McCarthy, J. (1979). Ascribing mental qualities to machines[18]. In Ringle, M., editor, *Philosophical Perspectives in Artificial Intelligence*. Harvester Press. Reprinted in [McCarthy, 1990].

[McCarthy, 1986] McCarthy, J. (1986). Applications of Circumscription to Formalizing Common Sense Knowledge[19]. *Artificial Intelligence*, 28:89–116. Reprinted in [McCarthy, 1990].

[16] http://www-formal.stanford.edu/jmc/

[17] http://www-formal.stanford.edu/jmc/mcc59.html

[18] http://www-formal.stanford.edu/jmc/ascribing.html

[19] http://www-formal.stanford.edu/jmc/applications.html

[McCarthy, 1989] McCarthy, J. (1989). Artificial Intelligence, Logic and Formalizing Common Sense[20]. In Thomason, R., editor, *Philosophical Logic and Artificial Intelligence*. Klüver Academic.

[McCarthy, 1990] McCarthy, J. (1990). *Formalization of common sense, papers by John McCarthy edited by V. Lifschitz*. Ablex.

[McCarthy, 1993] McCarthy, J. (1993). Notes on Formalizing Context[21]. In *IJCAI-93*. Available on http://www-formal.stanford.edu/jmc/.

[McCarthy, 1995] McCarthy, J. (1995). Situation Calculus with Concurrent Events and Narrative[22]. Contents subject to change. Reference will remain.

[McCarthy, 1996] McCarthy, J. (1996). Making Robots Conscious of their Mental States[23]. In Muggleton, S., editor, *Machine Intelligence 15*. Oxford University Press.

[McCarthy and Hayes, 1969] McCarthy, J. and Hayes, P. J. (1969). Some Philosophical Problems from the Standpoint of Artificial Intelligence[24]. In Meltzer, B. and Michie, D., editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press.

[Newell, 1981] Newell, A. (1981). The knowledge level. *AI Magazine*, 2(2):1–20. Originally delivered as the Presidential Address, American Association for Artificial Intelligence, AAAI80, Stanford, CA, August 1980.

[Newell, 1993] Newell, A. (1993). Reflections on the knowledge level. *Artificial Intelligence*, 59(1-2):31–38.

[Newell and Simon, 1972] Newell, A. and Simon, H. A. (1972). *Human Problem Solving*. Prentice–Hall, Englewood Cliffs, NJ.

[Turing, 1950] Turing, A. (1950). Computing machinery and intelligence. *Mind*.

[Turing, 1947] Turing, A. M. (1947). Lecture to the london mathematical society. In *The Collected Works of A. M. Turing*, volume Mechanical Intelligence. North-Holland. This was apparently the first public introduction of AI, typescript in the King's College archive, the book is 1992.

[20] http://www-formal.stanford.edu/jmc/ailogic.html
[21] http://www-formal.stanford.edu/jmc/context.html
[22] http://www-formal.stanford.edu/jmc/narrative.html
[23] http://www-formal.stanford.edu/jmc/consciousness.html
[24] http://www-formal.stanford.edu/jmc/mcchay69.html