
President's Message

WE NEED BETTER STANDARDS FOR AI RESEARCH

The state of the art in any science includes the criteria for evaluating research. Like every other aspect of the science, it has to be developed. The criteria for evaluating AI research are not in very good shape. I had intended to produce four presidential messages during my term but have managed only two, because this one has proved so difficult to write. It kept threatening to grow into a paper rather than a mere expression of opinion which is all I now know enough to write.

If we had better standards for evaluating research results in AI the field would progress faster.

One problem we have yet to overcome might be called the "Look, ma, no hands" syndrome. A paper reports that a computer has been programmed to do what no computer program has previously done, and that constitutes the report. How science has been advanced by this work or other people are aided in their work may be unapparent.

Some people put the problem in moral terms and accuse others of trying to fool the funding agencies and the public. However, there is no reason to suppose that people in AI are less motivated than other scientists to do good work. Indeed I have no information that the average quality of work in AI is less than that in other fields. In my previous message I grumbled about there being insufficient basic research, but one of the reasons for this is the difficulty of evaluating whether a piece of research has made basic progress.

It seems that evaluation should be based on the kind of advance the research purports to be. I haven't been able to develop a complete set of criteria, but here are some considerations.

1. Suppose the research constitutes making the computer solve a problem that hasn't previously been solved by computer.

Let us suppose that there are no theoretical arguments that the methods are adequate for a class of problems but merely a program that performs impressively on certain sample problems together with some explanation of how the program works.

This is a difficult kind of research to explain adequately. The reader will not easily be able to assure himself that the program is not overly specialized to the particular example problems that have been used in developing the program. It has often turned out that other researchers have not been able to learn much from the paper. Sometimes a topic is so intractable that this is the best that can be done, but maybe this means that the topic is too intractable for the present state of the art.

2. A better result occurs when a previously unidentified intellectual mechanism is described and shown to be either necessary or sufficient for some class of problems.

An example is the alpha-beta heuristic for game playing. Humans use it, but it wasn't identified by the writers of the first chess programs. It doesn't constitute a game playing program, but it seems clearly necessary, because without it, the number of positions that have to be examined is sometimes the square of the number when it is used.

3. Experimental work should be repeatable.

In the older experimental sciences, *e.g.* physics and biology, it is customary to repeat previous experiments in order to verify that a phenomenon claimed to exist really does or to verify a claimed value of an experimentally determined constant. The referees are supposed to be sure that papers describing experimental work contain enough of the right details so that this can be done.

Perhaps the most typical problem concerns a piece of experimental AI research, say a PhD thesis. The general class of problems that the researcher would like to attack is described, followed by a description of his program and followed by a description of the results obtained on his sample problems. Often there is only one sample problem. The class of problems which it is claimed the program or the methods it embodies will solve is often not stated. The reader is free to suspect that the program has been tuned so that it will solve the specific example described in the paper and that the author doesn't even know whether it will solve any others.

If we aspire to testable and repeatable work in AI, then journal authors and referees should require a statement of the generality of the program. The referee should be able to try out the program if language, hardware and communication facilities permit. Moreover, the methods should be described well enough so that someone skilled in the art can embody them in a program of his own and test whether they are adequate for the claimed class of programs.

Repetition of other people's experiments should be as normal in AI as it is in the other experimental sciences. On the whole, it should be easier in AI, because more-or-less standard hardware and programming languages can be used. Perhaps this is a good apprentice task for beginning graduate students or people coming into the AI field from the outside. Students and other newcomers will take pleasure in trying to find a simple example that the program is supposed to solve but doesn't.

Stating the generality of piece of work is likely to be difficult in many cases. It is best done after the program has solved the example problems, because the researcher can

then understand what compromises he has had to make with generality in order to make the program do his examples. He is most likely to make the necessary effort if he knows that some smart student is likely to look for counterexamples to his

4. *We also need criteria for formalizations.*

Logic based approaches to AI require that general facts about the common sense world be expressed in languages of logic and that reasoning principles (including non-monotonic principles) be stated that permit determining what a robot should do given its goals (stated in sentences), the general facts and the facts of the particular situation. The major criteria for judging the success of the formalization of such facts are generality and epistemological adequacy. Generality is partly a property of the language, and in the case of a first order language, this means the collection of predicate and function symbols. The original set of predicates and functions should not have to be revised when extensions are wanted. It is also partly a property of the set of axioms. They too should be extendable rather than having to be changed. The recent development of non-monotonic formalisms should make this easier.

The author of a paper proposing logical formalisms should state, if he can, how general they are. The referee and subsequent critics should try to verify that this is achieved.

Epistemological adequacy refers to the ability to express the facts that a person or robot in that information system is likely to be able to know and need to know.

5. *The criteria for evaluating methods that purport to reduce search are perhaps better established than in other fields.*

Taking my own experience with game playing programs in the late 1950s and early 1960s, it was possible to demonstrate how much alpha-beta, the killer heuristic, and various principles for move ordering reduce search.

6. *On the other hand, the evaluation of programs that purport to understand natural language is worse off.*

People often simply don't believe other people's claims to generality.

In this area I can offer two challenge problems. First, I can provide the vocabulary (sorted alphabetically) of a certain news story and the vocabulary of a set of questions about it. The computational linguistic system builder can then build into his system the ability to "understand" stories

and questions involving this vocabulary. When he is ready, I will further provide the story and the questions. He can take the questions in natural language or he can translate them into suitable input for his system. The limitations of the system should be described in advance. We will then see what questions are successfully answered and to what extent the author of the system understood its limitations.

The second problem involves building a system that can obtain information from databases that purport to interact with their users in English. Again the vocabulary is given in advance, and the system builder tunes his system for the vocabulary. It is then tested as to whether it can answer the questions by interacting with the database. For example, Lawrence Berkeley Laboratory has (or had) a database of 1970 census data. It would be interesting to know whether a program could be written that could determine the population of Palo Alto interacting with the interface this database presents to naive human users.

I think both of these problems are quite hard, and whatever groups could perform reasonably on them would deserve a lot of credit. Perhaps this would be a good subject for a prize—either awarded by the AAAI or someone else.

However, such challenge problems are no substitute for scientific criteria for evaluating research in natural language understanding.

7. *Likewise the Turing test, while a challenge problem for AI, is not a scientific criterion for AI research.*

The Turing test, suitably qualified, would be a fine sufficient criterion for convincing skeptics that human level AI had finally been achieved. However, we need criteria for evaluating more modest claims that a particular intellectual mechanism has been identified and found to be necessary or sufficient for some class of problems.

Incidentally, even as a sufficient condition, the Turing test requires qualification. The ability to imitate a human must stand up under challenge from a person advised by someone who knows how the program works. Otherwise, we are in the situation of someone watching a stage magician. We can't figure out the trick, but there must be one. *A fortiori*, looking at dialogs and figuring out which one is with a machine isn't adequate.

—John McCarthy
Department of Computer Science
Stanford University