END
DATE
FILMED
8-84
DTIC

MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

LUX ET VERITAS

The Explanation Game
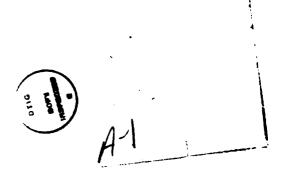
Roger C. Schank

YALEU/CSD/RR#307

March 1984

APR 3 0 1984

A

YALE UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE

The Explanation Game

Roger C. Schank

YALEU/CSD/RR#307

March 1984

A-1

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1 REPORT NUMBER #307	2. GOVT ACCESSION NO. AD-A140681	3. RECIPIENT'S CATALOG NUMBER
4 TITLE (and Subtitle) THE EXPLANATION GAME		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Roger C. Schank		8. CONTRACT OR GRANT NUMBER(s) N00014-75-C-1111 N00014-82-K-0149
9. PERFORMING ORGANIZATION NAME AND ADDRESS Yale University, Department of Computer Science 10 Hillhouse Avenue New Haven, CT 06520		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Advanced Research Projects Agency 1400 Wilson Boulevard Arlington, VA 22209		12. REPORT DATE March 1984
		13. NUMBER OF PAGES 26
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Office of Naval Research Information Systems Program Arlington, VA 22217		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16 DISTRIBUTION STATEMENT (of this Report)

Distribution of this report is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, If different from Report)

18 SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

learning
explanation
generalization
induction

20 ABSTRACT (Continue on reverse side if necessary and identify by block number)

In 1952 Alfred Turing proposed a test which he believed might serve as a touchstone for Artificial Intelligence researchers in their efforts to model understanding by computer. The essence of this test, which Turing called the Imitation Game, was to challenge a person to distinguish between a computer program and another human being by putting questions to each via teletype. Much has been made of this test in subsequent discussions concerning the ultimate success or failure of Artificial Intelligence (AI). Unfortunately,

the requirements of Turing's test are so rigorous that no program in existence today is close to succeeding, and none is likely to be for along time to come. As a result, such discussions tend toward metaphysics and away from practical scientific concerns. In this paper we examine the question of how to judge the success of an AI program from a research's point of view. At the heart of this discussion is the observation that understanding is not a unitary phenomenon. There are many levels of at which a program could be said to understand something, and if we are to evaluate the progress of AI in the near term we must be sensitive to the possibility that programs can understand at a somewhat lower level than we normally associate with humans. We therefore propose that AI programs be judged by the extent to which they can explain the decisions they make, and by the form which these explanations take. We present a classification of types of explanatory ability, and show a program's ability to explain its actions at a given level can give insight into the extent to which the program can be said to understand.

Although some may find it anathema to use the names of Turing and Weizenbaum in the same sentence, they each addressed the same issue. Turing (1950) asked the question, Can a Machine Think? in a paper that has become a classic. Weizenbaum (1970's) wrote a book that addressed a slightly different question, namely, can a computer really understand? These are two questions that will not go away as long as there are people working on creating intelligent machines, regardless of the success of that work. It seems that one way or the other, there will always be those who are certain of their answer to these questions, in principle, regardless of the particulars of the current research in progress.

After all, neither Turing's answer, wich was yes after he reworded the question, nor Weizenbaum's, which was a flat no, in any way depended upon the actual case with respect to computer programs extant at the time of the writing of their respective papers. Questions of this level of significance seem to engender gut-level responses in laymen, and philosophical responses in scholars. In this paper, I would like to present the third side of that argument, the researcher's side. I will discuss both Turing's and Weizenbaum's point of view, both of which I believe to be in error. I will then go on to give some perspective on the nature of understanding and what we mean by that word. I will conclude by attempting to tie together our notions of understanding and the nature of the kind of test a machine would have to pass in order to pronounce it cognizant.

-- OFFICIAL DISTIRUBTION LIST --

Defense Documentation Center Cameron Station Alexandria, Virginia 22314	12 copies
Office of Naval Research Information Systems Program Code 437 Arlington, Virginia 22217	2 copies
Dr. Judith Daly Advanced Research Projects Agency Cybernetics Technology Office 1400 Wilson Boulevard Arlington, Virginia 22209	3 copies
Office of Naval Research Branch Office - Boston 495 Summer Street Boston, Massachusetts 02210	1 copy
Office of Naval Research Branch Office - Chicago 536 South Clark Street Chicago, Illinois 60615	1 copy
Office of Naval Research Branch Office - Pasadena 1030 East Green Street Pasadena, California 91106	1 copy
Mr. Steven Wong New York Area Office 715 Broadway - 5th Floor New York, New York 10003	1 copy
Naval Research Laboratory Technical Information Division Code 2627 Washington, D.C. 20375	6 copies
Dr. A.L. Slafkosky Commandant of the Marine Corps Code RD-1 Washington, D.C. 20380	1 copy
Office of Naval Research Code 455 Arlington, Virginia 22217	1 copy
Office of Naval Research Code 458 Arlington, Virginia 22217	1 copy

Naval Electronics Laboratory Center   1 copy
Advanced Software Technology Division
Code 5200
San Diego, California   92152

Mr. E.H. Gleissner   1 copy
Naval Ship Research and Development
Computation and Mathematics Department
Bethesda, Maryland   20084

Captain Grace M. Hopper, USNR   1 copy
Naval Data Automation Command, Code OOH
Washington Navy Yard
Washington, D.C.   20374

Dr. Robert Engelmore   2 copies
Advanced Research Project Agency
Information Processing Techniques
1400 Wilson Boulevard
Arlington, Virginia   22209

Professor Omar Wing   1 copy
Columbia University in the City of New York
Department of Electrical Engineering and
Computer Science
New York, New York   10027

Office of Naval Research   1 copy
Assistant Chief for Technology
Code 200
Arlington, Virginia   22217

Computer Systems Management, Inc.   5 copies
1300 Wilson Boulevard, Suite 102
Arlington, Virginia 22209

Ms. Robin Dillard   1 copy
Naval Ocean Systems Center
C2 Information Processing Branch (Code 8242)
271 Catalina Boulevard
San Diego, California 92152

Dr. William Woods   1 copy
BBN
50 Moulton Street
Cambridge, MA 02138

Professor Van Dam   1 copy
Dept. of Computer Science
Brown University
Providence, RI 02912

Professor Eugene Charniak   1 copy
Dept. of Computer Science
Brown University
Providence, RI 02912

Professor Robert Wilensky   1 copy
Univ. of California
Elec. Engr. and Computer Science
Berkeley, CA 94707

Professor Allen Newell   1 copy
Dept. of Computer Science
Carnegie-Mellon University
Schenley Park
Pittsburgh, PA 15213

Professor David Waltz   1 copy
Univ. of Ill at Urbana-Champaign
Coordinated Science Lab
Urbana, IL 61801

Professor Patrick Winston   1 copy
MIT
545 Technology Square
Cambridge, MA 02139

Professor Marvin Minsky   1 copy
MIT
545 Technology Square
Cambridge, MA 02139

Professor Negroponte   1 copy
MIT
545 Technology Square
Cambridge, MA 02139

Professor Jerome Feldman   1 copy
Univ. of Rochester
Dept. of Computer Science
Rochester, NY 14627

Dr. Nils Nilsson   1 copy
Stanford Research Institute
Menlo Park, CA 94025

Dr. Alan Meyrowitz   1 copy
Office of Naval Research
Code 437
800 N. Quincy Street
Arlington, VA 22217

Dr. Edward Shortliffe   1 copy
Stanford University
MYCIN Project TC-117
Stanford Univ. Medical Center
Stanford, CA 94305

Dr. Douglas Lenat   1 copy
Stanford University
Computer Science Department
Stanford, CA 94305

Dr. M.C. Harrison   1 copy
Courant Institute Mathematical Science
New York University
New York, NY 10012

Dr. Morgan   1 copy
University of Pennsylvania
Dept. of Computer Science & Info. Sci.
Philadelphia, PA 19104

Mr. Fred M. Griffee   1 copy
Technical Advisor C3 Division
Marine Corps Development
   and Education Command
Quantico, VA 22134

Abstract

In 1952 Alfred Turing proposed a test which he believed might serve as a touchstone for Artificial Intelligence researchers in their efforts to model understanding by computer. The essence of this test, which Turing called the *Imitation Game*, was to challenge a person to distinguish between a computer program and another human being by putting questions to each via teletype. Much has been made of this test in subsequent discussions concerning the ultimate success or failure of Artificial Intelligence (AI). Unfortunately, the requirements of Turing's test are so rigorous that no program in existence today is close to succeeding, and none is likely to be for a long time to come. As a result, such discussions tend toward metaphysics and away from practical scientific concerns. In this paper we examine the question of how to judge the success of an AI program from a *researcher's* point of view. At the heart of this discussion is the observation that understanding is not a unitary phenomenon. There are many levels of at which a program could be said to understand something, and if we are to evaluate the progress of AI in the near term we must be sensitive to the possibility that programs can understand at a somewhat lower level than we normally associate with humans. We therefore propose that AI programs be judged by the extent to which they can explain the decisions they make, and by the form which these explanations take. We present a classification of types of explanatory ability, and show a program's ability to explain its actions at a given level can give insight into the extent to which the program can be said to understand.

Although some may find it anathema to use the names of Turing and Weizenbaum in the same sentence, they each addressed the same issue. Turing (1950) asked the question, *Can a Machine Think?* in a paper that has become a classic. Weizenbaum (1976) wrote a book that addressed a slightly different question, namely, *can a computer really understand?* These are two questions that will not go away as long as there are people working on creating intelligent machines, regardless of the success of that work. It seems that one way or the other, there will always be those who are certain of their answer to these questions, in principle, regardless of the particulars of the current research in progress.

After all, neither Turing's answer, which was *yes* after he reworded the question, nor Weizenbaum's, which was a flat *no*, in any way depended upon the actual case with respect to computer programs extant at the time of the writing of their respective papers. Questions of this level of significance seem to engender gut-level responses in laymen, and philosophical responses in scholars. In this paper, I would like to present the third side of that argument, the *researcher's* side. I will discuss both Turing's and Weizenbaum's point of view, both of which I believe to be in error. I will then go on to give some perspective on the nature of understanding and what we mean by that word. I will conclude by attempting to tie together our notions of understanding and the nature of the kind of test a machine would have to pass in order to pronounce it cognizant.

The Turing Test

In addressing the question *can a machine think*, Turing took a strict empiricist line. The question, he said, was meaningless. Thought, after all, is not something which can be defined concretely enough to test for. He might have left it at that were it not for the fact that the field of Artificial Intelligence seemed to be defined as the attempt to construct machines for which the answer to this meaningless question would be *yes*. Rather than leave AI to wander forever in search of this holy grail, Turing proposed that the success of the field be tied to a more down-to-earth question, the passing of a specific test. The test he proposed was called the *Imitation Game*. The test is to see if a person. communicating via teletype, can distinguish between a man imitating a woman and a computer imitating a man imitating a woman. If the person could not tell the difference, the computer passed the test.

Though both the actual workings of the test and Turing's reasons for proposing it are frequently misunderstood, the idea of a test in general is both compelling and easy to grasp. Thus it has become a popular conception of how to judge the ability of a machine to understand. Quite naturally, two lines of argument have developed concerning the test. One is the issue of whether a machine will ever pass it. The other is the issue of whether passing the test means that a machine is thinking, since building a machine that thinks, or understand, remains the point of Artificial Intelligence for researchers and laymen alike, despite Turing's efforts.

Turing went on to argue that machines would eventually pass his test. Since there were no programs around when he wrote that even showed signs of being able to accomplish this feat, he was forced to argue from nothing other than beliefs, despite his seeming stubborn empiricism. And since in the thirty years since the paper was written no program has passed the Turing test[1] subsequent discussion of the issue has been similarly divorced from the actual accomplishments of current AI. Thus, instead of moving the debate to solid scientific ground, Turing's proposal had the effect of furthering the distance between the question of whether machines can think and the consideration of the actual state of the science of AI. Because of this, no achievement of AI, and no program that works well, seems to impinge on this debate at all. I believe that it is time to look again at the twin questions of the criteria for success of AI programs, and the nature of what is meant by understanding.

---

[1] Colby (1973) claimed that he had a program which had in some sense passed the test, but this program actually simulated a deeply psychotic person, so it isn't exactly what Turing had in mind.

On Men and Women

One of the interesting facets of Turing's Imitation Game is the involvement of the task of imitating a woman. This is interesting in part because it is hard to see what difference it makes to the game. One would expect the game to come down to the question of whether the computer can simulate the man, period. It is not clear why a computer would perform any differently on the task of simulating a man simulating a woman than on the task of simulating a man being himself. Perhaps the fact that both the computer and the man would be forced to fabricate their answers makes a difference, although this does not seem likely to be an important issue. Certainly, the test has popularly been taken to be simply that of a machine trying to simulate a person on a teletype.

The formulation of the game probably has to do with the fact that Turing's original conception of the game was simply a man trying to simulate a woman over a teletype (and being compared to an actual specimen by the subject). Perhaps Turing's test for the computer was formulated as it was due to some strange sense of symmetry on Turing's part. Regardless of this, however, Turing certainly was basing his game on an analogy between the act of a man trying to imitate a woman and the act of a computer trying to imitate a person. Though Turing's test does not really depend on men and women being discernably different in this test, the analogy is interesting as an illustration of what it means to understand. Not the least of the aspects of thinking which the Turing test bears on is the ability to understand a human being. Presumably, this is the ability which lies at the heart of being able to imitate someone successfully, and making a program understand people to a reasonable degree would be the central problem in designing a program to play the Imitation Game.

To see how the analogy is relevant, let us consider the question of whether men can really understand women, (or alternatively, whether women can really understand men). It is common enough in everyday experience, for men and women to both exclaim that they really do not understand their opposite number. What can they mean by this? And, most importantly, how is what they mean by it related to the problem of determining whether computers can understand?

When the claim is made that men and women are really quite different mentally (not physically), what is presumably meant is that they have different beliefs, different methods of processing information, different styles of reasoning, different value systems, and so on. (It is not my point here to comment on the validity of these assertions, I am simply attempting to use the principle used by such assertions in my argument. These same arguments can be made about different ethnic groups, cultures, nations and so on, but men and women will do for now.)

The claim that I assume is not being made by such assertions, is that men and women have different physical instantiations of their mental processes. (Of course, it is possible that men and women do have brains that differ physically in important respects, but that would be irrelevant for this argument as I shall show.) So, what is it that makes it seem to both men and women that they have difficulties understanding each other? The answer is that, despite possibly physically identical brains, understanding involves empathy. That is, it is easier to understand someone who has had similar experiences, and who, because of those experiences, has developed similar values, beliefs, inference structures, MOPs, scripts, goals, and whatever other structures for holding information might exist in the mind.

Let's consider this another way. Understanding, we claim, consists of processing incoming experiences in terms of the cognitive apparatus one has available. This cognitive apparatus has a physical instantiation (the brain or the hardware of the computer) and a mental instantiation (the mind or the software of the computer). When an episode is being processed, a person brings to bear the totality of his cognitive apparatus to attempt to understand it. What this means in practice is that people understand things in terms of their particular memories and experiences. Specifically, this means that people who have different goals, beliefs, expectations, and general life styles, will understand identical episodes quite differently.

In essence then, we are saying that no two people understand in exactly the same way or with the same result. The more different people are, the more different will be their perception of their experiences. And, it follows that when people share certain dimensions of experience, they will tend to perceive experiences along those dimensions similarly. Thus, men tend to understand certain classes of experiences differently than women.

It is unlikely that an experience that in no way bears upon one's sex will be understood differently by men and women. Recall that the assumption here is that the base line cognitive apparatus is the same regardless of sex. But, we are claiming that any experience that does relate to the sex of the observer in some way, even obscurely, will be processed differently. This can involve obvious issues, such as the observation of an argument between a man and a women. There, we might expect a man to observe the episode from the point of view of the man and the women to observe it from the point of view of the woman. In addition such identification with different characters in a situation can extend to observations of situations where the feuding characters are of the same sex, but one displays attributes more traditionally *male* and the other similarly displays *female* behavior. Identification, and thus perception, can thus be altered by one's understanding of the goals, beliefs, or attitudes underlying or perceived to underlie the

behavior of the characters in an episode one is observing.

Thus, for example, one's perception of the validity and purpose behind a war can be altered by whether one is the mother of a son who is about to be drafted, or whether one previously fought in a war and found it an ennobling experience. In general, one's sense of what is important in life affects every aspect of one's understanding of events.

The claim then, is that men and women, as examples of one division of human beings, do not, and really cannot, understand each other, in the deepest possible sense of that word. The same argument can be put forward, with more or less success, depending upon the issue under consideration, with respect to Arabs and Israelis, intellectuals and blue collar workers, and so on. In each of these cases, differing values can cause differing perceptions of the world.

Computers and People

The issue of understanding a person specifically relates to Weizenbaum's answer. His claim was that computers would never really understand, in the sense that a computer, because of its lack of experience in human affairs of the heart, would be unable to truly understand a *shy young man's desperate longing for love* as expressed by his dinner invitation to a woman.

The problem is that understanding is not such a simple affair. I, for one, feel comfortable with admitting that no computer would ever understand this shy young man's desperate longing, in the fullest sense of understanding. On the other hand, it seems obvious to me that the computer could still understand a story involving a dinner date between this shy young man and his object of affection. It is possible to hold such seemingly contradictory views because understanding itself is not such a unitary affair. There are a great many kinds of understanding. I should like to consider some of them here.

The Nature of Understanding

What I would like to do is discuss a spectrum of understanding that will allow us to more sensibly consider the issue. Let's consider some of the points on that spectrum. At the far end of the spectrum, we have what I shall call COMPLETE EMPATHY. This is the kind of understanding that might obtain between twins, very close brothers, very old friends that know each other's every move and motivation, and other such combinations of people that rarely are found in the world.

At the opposite end of the spectrum we have the barest form of understanding, which I shall call MAKING SENSE. This is the point where events that occur in the world can be interpreted by the understander in terms of a coherent (although probably incomplete) picture of how those

events came to pass, despite the understanders complete lack of empathy with the actors in those events. Between these two poles lies a continuum of understanding differentiated by the amount of personal experience the understander is able to bring to bear on the events he is perceiving. Loosely, we might describe the end points as situations where the understander says to himself *yes, I see what is going on here, it makes some sense to me,* and, on the other hand, his saying *my god, that's exactly what I would have done, I know precisely how you feel.*

In our recent research (for example, Schank and Riesbeck, 1981 and Schank, 1982) we have been concerned with the nature of understanding because we are trying to get computers to read and process stories. In the course of that research, we have considered various human situations that we wished to model. Our attempt to model those situations has resulted in our building various *knowledge structures* that attempt to characterize the knowledge of various situations that people have. One well-known such knowledge structure is the restaurant script (Schank and Abelson, 1977). It was our use of such scripts to attempt to understand restaurant stories, in fact, that prompted Weizenbaum's criticism about understanding *love in a restaurant.*

The basic hypothesis behind our work is the notion that in attempting to understand we are attempting to relate our new experiences to our prior experiences by utilizing knowledge structures that organize those prior experiences. Consider, for example, the following situation. Imagine yourself going to a Burger King under the circumstances where you have been to McDonald's on numerous occasions but have never before been to Burger King. You are confronted with a new situation which you must attempt to "understand". We can say that a person has understood such an experience (i.e. he understands Burger King in the sense of being able to operate in it) when he says *oh I see, Burger King is just like McDonald's.* To put this another way, we might expect that at some point during your Burger King trip you might be *reminded* of McDonald's. The point here is that understanding, on any part of the spectrum of understanding mentioned above, means being reminded of the closest prior experienced phenomenon in memory and being able to use the expectations generated by that reminding to help in processing the current experience.

When we are reminded of some event or experience in the course of undergoing a different experience, this reminding behavior is not random. We are reminded of this experience because the structures we are using to process this new experience are the same structures we are using to organize memory. Thus, we cannot help but pass through the old memories while processing a new input. There are an extremely large number of such high level memory structures. Finding the right one, or the right set, of these, (that is, those that are most specific to the experience at

hand) is what we mean by understanding.

The Spectrum of Understanding

Given the sense of the nature of understanding above it seems clear that one reason why COMPLETE EMPATHY might exist would be that so many shared experiences between the individuals involved would have caused very similar memory structures to have been created. The consequence of this is that, given a set of similar goals and beliefs, new episodes would be processed in the same way. The above caveat is very important. Similar experiences, but different goals and beliefs, would still result in differing perceptions of the events, or to put it another way, in a lack of COMPLETE EMPATHY in understanding of each other's actions.

The point to be made about the understanding spectrum then, is that the more completely that goals, beliefs, prior experiences and memories are shared, the more complete understanding can take place. On the opposite end of the spectrum, MAKING SENSE involves finding out what events took place and relating them to a perception of the world that may be quite different from that in the mind of the actor in those events.

To make all this more clear, let us now consider what may well be a midpoint on the understanding spectrum. We discussed earlier the problem of men and women understanding each other, in general. The point there was that despite a cognitive apparatus that was identical, something was preventing complete understanding. This mid-point I will label, COGNITIVE UNDERSTANDING. By this I mean that while a man may be able to build an accurate model of a given woman (his wife for example), he may still not really understand what her motivations, fears, needs, and so on are. That is he lacks COMPLETE EMPATHY with her, but he still understands a great deal about her. To claim that he doesn't understand her can only mean understanding in its deepest sense. Certainly by any measure of understanding less than COMPLETE EMPATHY he could rightly claim to understand what she does and not be accused by Weizenbaum of failing the understanding test.

Thus, there are obviously many different kinds of understanding. Where do computers fit in? The claim I want to make is that, given a spectrum as we have described:

MAKING SENSE----------COGNITIVE UNDERSTANDING-----COMPLETE EMPATHY

today's work on computer understanding can only reasonably claim the left half of the spectrum as its proper domain. Weizenbaum has a legitimate argument in claiming that

computers will never understand, if what he means by that is that they would be unlikely to understand much at the right hand side of the spectrum. In other words, it is an easy concession to make to state that computers do not actually feel the same way as people, do not have the same goals as people, and are in fact, not people. Given that most people fail to understand each other at points on the right hand side of this spectrum, it seems no great calumny to admit to the likely failure of computers to achieve this level of understanding. Computers are certainly unlikely to do better than women at understanding men, blue collar workers at understanding academics, Arabs at understanding Israelis, or any of the corresponding vice versas.

But, on the other hand, computers can and do, share some experiences with people. They can, in principle, read newspapers and build models of the situations they read about almost as well as people do. I do not feel in reading a news story about terrorism, that, just because the terrorist and I share the feature of being human (and even perhaps the same age and sex), that I really understand that terrorist's actions. And, although being capable of human feelings may facilitate my understanding of the story vis-a-vis the computer, the computer's long history of reading and remembering terrorism stories (as in Lebowitz, 1980) can equally well be said to facilitate its understanding of such stories.

Thus, I argue, Weizenbaum was in error when he claimed that computers cannot understand, if he was referring to the left hand side of the above spectrum. On the assumption that he was referring to the COMPLETE EMPATHY side, the argument there is more complex. No computer scientist should argue that people and computers will ever have complete empathy for each other since they will have had such different experiences. Computers are not likely to experience the need for sex, love, food, possessions, and so on, and are thus unlikely to see the world exactly the way people see it. On the other hand, intelligent machines ought naturally be able to find some common ground of shared experience with the people with whom they interact. Thus, Weizenbaum is correct when he points out that shared experience is critical in understanding. He just takes this notion too far. Computers cannot, in principle, completely understand people, any more than men can completely understand women or vice versa. But, that point of COMPLETE EMPATHY is just that, an end point on a continuum. There is a lot of space in between the end points.

A more interesting question here is what computer understanding at any of these levels would tell us about either human understanding or computers. To put this another way, how different are these three points on the spectrum in terms of what processes would be necessary in order to construct machines that met those criteria? What would a machine have to do in order to

achieve the ability to MAKE SENSE, COGNITIVELY UNDERSTAND, or show COMPLETE EMPATHY?

The Revised Turing Test

Turing's Imitation Game has not left everyone in AI thrilled by the prospects of having to meet its criteria as a measure of success. Colby (1973) argued that his paranoid simulation program did indeed pass Turing's test. He found that psychiatrists who were presented with output from PARRY and output in a similar form from an actual human patient, were unable to effectively distinguish between them. This *passing of the Turing test* that Colby claimed failed to convince AI people that Colby's program *understood* or *thought*, nor should it have done so. Despite whatever validity Colby's program might have as a model of paranoia, because it was by definition simulating abnormal behavior, not normal human thought, the failure of the experts to distinguish between imitations and the real thing should not be taken as much more than a statement of the competence of the experts. In fact, in this case, even that is going too far. The Imitation Game was a particularly unfair test since a psychiatrist's choices upon facing a non-normal patient are not that extensive. PARRY was not brain-damaged or schizophrenic, so paranoid was, given the presence of a few paranoid signs, a reasonable choice. What Colby seems to have done is effectively simulated various aspects of the output of a paranoid, and it was possible to get by with this because in the context of dealing with what was supposedly highly abnormal human behavior, virtually any failure of the program to perform like a human being could be excused as merely another display of abnormal behavior by the "patient".

The case of Parry brings up an interesting point about judging the ability to "understand" of a system whose limited task saves it from having to exhibit intelligent behavior over the whole range of tasks on which humans can perform. While it can be argued that a program which somehow simulated the output of a normal human being effectively in any context in which it was asked to do so might have a claim to being called intelligent regardless of the manner in which the simulation is performed (i.e., whether or not it bears any relationship to the way people think), the argument is not so convincing for a program which only performs convincingly in a narrow context. It is always possible to isolate a subset of intelligent behavior which can be simulated by means clearly unrelated to what we think of as intelligence; chess-playing programs which operate by brute-force search techniques are the classic example of this. The problem is that no program in AI has yet gone beyond a narrow domain of performance. Therefore, if we want to have metrics for the progress of AI towards intelligent machines, we will have to employ tests which look at something other than simple external behavior.

Thus the best test of an effective understanding system, if I may use that word, is not the

realism of the output it produces, as Turing would have it, but rather the validity of the method by which that output is produced. Unfortunately, we cannot create a test that depends upon the evaluation of methods. We cannot do this for two reasons, one practical and one philosophical. The practical reason is that it is difficult to cut open either the machine or the human to see what he is doing. Examination of hardware doesn't tell us much anyway. From a software point of view, people are extremely hard to test in a controlled experiment, given the wide range of their possible experiences prior to that experiment. And, programs are hard to evaluate with respect to establishing exactly what claims they are making.

The philosophical point is that we do not examine the innards of people in order to establish that they are understanding. It might seem that we grant them an ability to understand based on their humanity. But, it oftens happens, in schools, offices, and other places that do such evaluations, that we do assess a human's ability to understand. It seems sensible therefore, that the methods that are good for the human are also good for the computer.

To see what I mean here, let's consider the spectrum of understanding again. For the sake of argument, assume that it is possible, in theory, to produce a program that meets the output requirements of each of the three levels of understanding noted above. To make this argument more concrete, I will list below some possible outputs for each of these points on the spectrum:

MAKING SENSE

input: news from the UPI wire
output: a summary of a newspaper story
    a translation of a speech into another language

COGNITIVE UNDERSTANDING

input: a set of stories about airplane crashes, complete with data
    about the airplanes and the circumstances
output: a conclusion about what may have caused
    the crash that turns out to be accurate

COMPLETE EMPATHY

input: I was very upset by your actions last night.
output: I thought you might have been, it was lot like the
    way you treated me last week.
input: But I meant you no harm.

output: Do you remember the way your father used to treat you
      on holidays when he made you call all your relatives?
      He meant no harm either.
input: I see what you mean.
output: I thought you might; there's no friend like an old friend.

Assuming that these input/output pairs are not entirely fanciful, I would now like to draw some conclusions about them that are reflective of my view of what a reasonable test should comprise. My conclusions are effectively summarized with the following words:

ACCURACY; SURPRISE; EMOTION

The claim I am making is, to the extent that output is an effective way of characterizing degree of understanding, (although that is to a very limited extent indeed, it may well be our only choice), we can judge the significance of that output in terms of its place on the understanding spectrum with respect to the following features:

The extent that that output accurately accomplishes a task that a competent human could do.

The extent that that output characterizes an especially original or important result that most humans could not easily accomplish.

The extent that that output effectively seems to replicate a real live human being whom someone is familiar with.

I would now like to propose a successor to the Imitation Game, which I will call the Explanation Game.

The Explanation Game

In the end, any system, human or mechanical, is judged on its output. We do not take apart a human to look at his insides in an effort to establish that his understanding mechanisms are of the right sort. Nor is it clear what the right sort of mechanisms are. We are faced with a dilemma then. We cannot use output to tell us if a system *really understands*. On the other hand, output is all we can reasonably expect to get.

To resolve this dilemma we must address the question of self-awareness. The issue of consciousness is a vast one and I do not mean to address it here. Rather, I wish to claim that the fundamental difference between a system that can produce reasonable output and one that meets the criterion that we mean by the label **understanding system** is that an understanding system should be able to explain its own actions. A system that not only does interesting things, but can explain why it did them by relating what it did to other episodes and circumstances in its world can be said to understand at any point on the understanding spectrum where that explanation is sufficient. This is not to say that a system that cannot explain cannot therefore understand.

The question is how an understanding system can show an outside observer that it has understood. Humans are happy to attribute understanding capabilities to everything from cats to dolphins. Somehow, however, machines fail to enjoy such largesse. So while we cannot claim that if you cannot explain you have not understood, we can claim that if you can explain you must have understood.

To make this more concrete, consider the three points on the spectrum that we have discussed. To satisfy understanding requirements at the MAKING SENSE level, consider SHRDLU, Winograd (1972). That program manipulated blocks in a blocks world by responding to English commands. One of the aspects of that program that was most impressive was that when asked why the program had performed any given action, the program could respond with a causal sequence of actions. It could say that it had done X to do Y in order to do Z and so on. At the end of its chain of reasoning, it had only the initial command given by the user. In that case, it would respond, *I did that because you asked me to.* This latter response became well-known and was psychologically very appealing. (For example, Restak (1979) used the phrase *Because you asked me to* as the title of a chapter in a popular book on the brain. That chapter was on various aspects of AI and only touched upon Winograd's work.) Although it was not put this way at the time, one of the reasons why Winograd's program was much appreciated was because it "understood" at the MAKING SENSE level on the understanding spectrum. It understood its world of blocks as well as a human would. Now, it would not have passed Turing's test, because it understood nothing but blocks. But, I claim, it should be recognized as having passed the Explanation Game test at the level of MAKING SENSE, the criterion of which is:

for MAKING SENSE: Any program that, in performing a set of actions that simulate human output in a task requiring it to MAKE SENSE of the input, that can effectively explain why it did what it did, as well as a person could explain those same actions, understands at the level of MAKING SENSE

Each point on the understanding spectrum has essentially the same requirements in the Explanation Game. For COGNITIVE UNDERSTANDING, the program must be able to explain why it came to the conclusions it did, what hypotheses it rejected and why, how previous experiences influenced it to come up with its hypotheses and so on. We do not let a human being come up with innovative ideas, generalizations, correlations and so on, unless he can explain himself. There is no reason why we should let a machine off more easily. The machine must be able to answer the question *How do you know?* to the satisfaction of an examiner in a school who would expect no less from his students on a similar accomplishment in a similar domain.

The last point we have presented on our spectrum, that of COMPLETE EMPATHY, has no easier test. Any system purporting to satisfy this level of understanding must satisfy its examiner in the same way that a human would satisfy him in a similar situation. I am claiming that this is not likely in its fullest sense, and that this improbability is what all the uproar is about with respect to the assessment of the possibilities for Artificial Intelligence by laymen. No machine will have undergone enough experiences, and reacted to them in the same way as you did, to satisfy you that it *really understands* you. One is extremely lucky if one meets one person in one's lifetime who satisfies that criterion.

It was suggested by Riesbeck (personal communication) that one way to make the distinction between passing the Explanation Game at each level is to use the example of a joke. A computer understander that simply understood the joke, in that it could explain what had happened, would be understanding at the level of MAKING SENSE. A program that actually found the joke funny, to the extent that it could explain what expectations had been violated, what other jokes it knew that were like it that it was reminded of, and so on, would be understanding at the level of COGNITIVE UNDERSTANDING. Finally, a program that belly laughed because of how that joke related to its own particular experiences and really expressed a point of view about life that the program was only now realizing, would have understood at the level of COMPLETE EMPATHY.

Are any of these three levels likely to eventuate for computers? I think is clear that AI has already passed the starting point on the left hand side of the spectrum and is heading towards COGNITIVE UNDERSTANDING. Can we pass that point?

For all points to the right of COGNITIVE UNDERSTANDING, the question of what explanations and discussions of one's behavior satisfy the understanding requirement is more one of personal feeling than of science. Artificial Intelligence should, and I predict will, be happy with achieving the simulation of COGNITIVE UNDERSTANDING. It is unlikely to satisfy very many examiners in the Explanation Game at the extreme right end of the spectrum. Nor should it. We are not trying to build people, only intelligent machines.

Levels of Explanation

The key question then is how we are to differentiate the various kinds of explanation that are required to demonstrate understanding be it human or computer understanding. In other words, what does it take to win the explanation game? How can we know if we have moved from left to right on the understanding spectrum?

In this section we shall consider how the test of explanation be made. The test should

hopefully not merely be some ad hoc device. Rather, the explanations we seek should already be inherent in any program that purports to understand. We shall also consider what explanation tells us about intelligence, which is the true subject matter of AI.

Bear in mind that explanation is different depending upon the point in the understanding spectrum that we are dealing with. At each point, some type of explanation is required. It is the depth and type of the explanation that varies. So, at the level of making sense, consider the following example. We expect, when hearing of an event that we wish to make sense of, to be able to put the pieces together in a whole. Thus, if we hear an account of an event, we listen to see if each element in the event relates to its parts. When it does not, we attempt to make inferences that tie together the individual elements. When we cannot make such inferences, either one of two things is the case. We may have a situation where there exists an appropriate knowledge structure that, if known, would tie together the seemingly unrelated pieces. If we do not know this structure, then we can ask about it, or attempt to discover it for ourselves. On the other hand, it may be the case that this structure does not exist. In that case, we usually say that what we heard didn't "make sense."

Example

The point here is that both computers and people require, for the most minimal level of understanding, a set of information that they can rely upon to tie together elements in an event that are connected but whose connection has not been explicitly stated. Thus, the first type of explanation it is important to recognize is COHERENCY EXPLANATIONS, that is the type of explanation that relies upon a store of knowledge to draw inferences that create connectivity in a text, scene, or plan.

Samples

The second type of explanation that we shall discuss, corresponding to the level of COGNITIVE UNDERSTANDING, is that of FAILURE EXPLANATIONS. The premise of COGNITIVE UNDERSTANDING is basically that an input, after being processed to see if all of its pieces are connected, must be processed in such a way as to relate it to previously stored experiences that a system has already processed. The process of doing this was outlined in Schank (1982). The premise there was that when an input fails to conform to expectations derived from prior experience, its differences are noted and stored. In order to achieve real insights into why an input did not match one's expectations, one must attempt to explain the failure. This explanation is then used as an index to that particular experience. When another experience fails an expectation and is explained the same way, people get reminded. Hence we must attempt to make this the case with machines. This reminding serves as the basis of

learning (as discussed in Schank, 1982). People learn by comparing experiences that differed from expectations in the same way, so as to enable the creation of a new set of expectations that capture the generalizations created by similar failures with similar explanations.

So, the first critical role of explanation is simply the tying together of events in such a way as to fill in the missing pieces so as to make sure that a smooth chain of causality exists. If this chain of causally-linked events can be created, then we can argue that a system has understood, at the MAKING SENSE level. The second level of explanation implies a deeper level of understanding. Explaining failures implies the ability to understand by relating a set of events to one's own personal experiences, since failures are by definition violations of expectations derived from one's own experience. Relating information being processed to prior information that was already processed is a qualitatively different kind of understanding. Explaining at that level is the basis of learning and thus the basis of a system that can be surprising in some sense. Thus, COGNITIVE UNDERSTANDING implies learning, and thus, in a limited sense some elements of creativity.

The third level, COMPLETE EMPATHY, also has connected to it a level of explanation. Often, when we hear somebody tell us about his daily problems, we respond with anecdotes based upon our own experiences. But, there is a deeper level of understanding of another individual that is possible, and that is one that is based on a fairly sophisticated model of that individual that we have built up over time. As we get to know someone, we can build up a model of that person that explains why he acts the way he does and predicts, to some degree, his future behavior. In a sense, the notion of COMPLETE EMPATHY that we have suggested is just the end point on a rather wide spectrum that encompasses understanding of this kind. The more we know about another human being, the more we are able to understand him. (There is that old saying that in order to criticize someone you should first walk a mile in his shoes.)

The third type of explanation is what I shall call CONTRIBUTORY EXPLANATION. Here, what we are seeking is an understanding of the reasons behind an action that someone takes. When Weizenbaum's ELIZA program responded to its "patient" that her boyfriend and her father and bullies were all intertwined, it was that response that made people believe that was something to whatever it was that Weizenbaum had done. People were impressed because there was a CONTRIBUTORY EXPLANATION. Now, in this case, what we really had was a gimmic that was in no way generalizable. But, if it had been done by a method that was based on a theory of how such connections can be made in understanding the rationale behind the behavior of a patient, then we would have been witness to an impressive piece of work.

Computer Examples

Consider the following computer programs. These programs, are actually being developed at my AI lab at Yale, but for my purposes here that is not important. Rather, I will talk about these programs without too much regard for their current status, to illustrate the way explanation can be incorporated into AI programs in practice. The programs I want to talk about have one thing in common: they attempt to model an expert in the way he reasons about his field of expertise. The programs are a chef and a football coach.

In the CHEF project (Hammond, 1983), the goal is to develop a system that can build original plans, in the form of executable recipes, to deal with new situations. In order to do this the system needs the ability to create new plans, but more importantly, it must have the ability to examine and explain the results of those plans in order to debug not only the individual plans but also the program's understanding of the world. One of its primary strategies is to produce a plan and then use the feedback from the results of running the plan to modify both its plan and its understanding of the domain. This approach is particularly useful in situations where many different rules can be applied to predict the outcome, but the results of the plan indicate which rules are applicable to the current situation.

In this example we are looking at the analysis of a situation in which a plan has been tried and we are attempting to explain it in order to correct it. The task presented was to build a strawberry souffle. The plan consisted of pulping fresh strawberries and adding them to an existing plan for vanilla souffle. The result is that the souffle fell. It is important to note that this is not a program log but is instead a trace of the model. As we enter, the souffle has fallen.

Why has the souffle fallen?

SOMETHING HAS INTERFERED WITH THE ABILITY OF THE EGG-WHITES
TO KEEP IT UP.

Have any new steps been added?

NO NEW STEPS WERE ADDED TO THE ORIGINAL PLAN.
ALL NEW STEPS ARE ON NEW INGREDIENTS.

What are properties of the new ingredients?

WHEN ADDED, THE TEXTURE OF THE STRAWBERRIES WAS THAT OF HEAVY LIQUID.
NO PHYSICAL EFFECTS ARE NORMALLY ASSOCIATED WITH STRAWBERRIES.

IN THE PAST A SOUFFLE FELL THAT HAD A LIGHT LIQUID ADDED TO IT.
ADDING LIQUID INTERFERES WITH THE ACTION OF THE EGG-WHITE.

THE LIQUID FROM THE STRAWBERRY INTERFERED WITH THE EGG-WHITE.

IN SITUATION
    SIDE-EFFECT OF PLAN BLOCKING PRE-CONDITIONS OF EXISTING PLAN.
      PROBLEM IS LEVEL OF INTENSITY.
     SIDE-EFFECT INCREASES EXISTING CONDITIONS.

    THE ADDITION OF THE STRAWBERRY TO GET THE TASTE HAS ADDED MORE
    LIQUID WHICH BLOCKS THE ABILITY OF THE EGG-WHITES TO KEEP
    THE SOUFFLE RISEN.

    THE EGG-WHITES CAN HANDLE SOME LIQUID BUT NOT MORE.

    OUR GENERAL PLANS ARE -

        FIND PLAN FOR FIRST GOAL THAT DOES NOT HAVE SAME
        SIDE EFFECTS.

        RECOVER FROM SIDE EFFECTS.

        COMPENSATE ORIGINAL PLAN TO OVERCOME SIDE-EFFECTS.

        CHANGE ORIGINAL PLAN TO ONE WITH DIFFERENT
        PRE-CONDITIONS.

    TRYING

        FIND PLAN FOR FIRST GOAL THAT DOES HAVE SAME
        SIDE EFFECTS.

           LOOKING FOR INGREDIENT THAT HAS TASTE OF STRAWBERRIES
           BUT IS NOT LIQUID.

           FOUND STRAWBERRY JAM.

           ALTERING PLAN TO USE STRAWBERRY JAM INSTEAD OF FRESH
           STRAWBERRIES.

        RECOVER FROM SIDE EFFECTS

           NO WAY TO REMOVE LIQUID AFTER A MIX OF LIQUID
           AND LIQUID.

        COMPENSATE ORIGINAL PLAN TO OVERCOME SIDE-EFFECTS.

           ORIGINAL PLAN INVOLVES BEATING EGG-WHITES.

           TRYING MORE BEATING -
             REJECTED BECAUSE EGGS ARE TO BE FLUFFY
             NOT STIFF.

```
            TRYING MORE EGG-WHITES

            ALTERING PLAN TO USE EXTRA EGG-WHITES

        CHANGE ORIGINAL PLAN TO ONE WITH DIFFERENT
        PRE-CONDITIONS.

            LOOKING FOR OTHER RISING AGENT.

                FOUND FLOUR USED IN CAKES.

            ALTERING PLAN TO ADD FLOUR.

    THREE ALTERNATIVES

            ALTERING PLAN TO USE STRAWBERRY JAM INSTEAD OF FRESH
            STRAWBERRIES.

            ALTERING PLAN TO USE EXTRA EGG-WHITES

            ALTERING PLAN TO ADD FLOUR.

    CHOOSING ALTERING PLAN TO USE EXTRA EGG-WHITES

            THIS PLAN DOES NOT ADD NEW INGREDIENTS.
            ALTERING PLAN TO ADD FLOUR DOES.

            THIS PLAN MATCHES GOAL TO HAVE FRESH
            STRAWBERRIES. ALTERING PLAN TO USE
            STRAWBERRY JAM INSTEAD OF FRESH STRAWBERRIES
            DOES NOT.
```

The CHEF program is dealing in FAILURE EXPLANATIONS. Its explanations will therefore serve not only the goal of displaying the program's understanding of the concepts it is manipulating to create recipes, but is also necessary to the program's functioning because the explanations it generates serve to suggest the plan modifications it should try.

The COACH program ultimately has the same goal as the CHEF program: creating novel plans for accomplishing its objectives, in this case winning football games. The underlying relationships between subgoals of the goal to win the game -- score, keep your opponent from scoring, use up/preserve the time on the clock, wear your opponent down, keep your team fresh, lull your opponent or surprise him, and so on -- are very complex. The first task of a system to create novel plans to fulfill this galaxy of complicated goals is to understand how those goals

relate to the execution of its plans, which in this domain consists of calling offensive plays. The current system calls plays from its library of known plays, so it basically functions as an *expert system* for play calling. Merely being able to call reasonable plays, however, is no indication that the program in any sense understands the goals it is pursuing or the relationships among them. For example, a system to call plays might simply be full of rules like the following hypothetical example:

```
IF DOWN > 2 AND
   DISTANCE-TO-FIRST-DOWN < 3 AND
   DISTANCE-TO-FIRST-DOWN > .5 AND
   (TIME > 2 OR OUR-SCORE >= OPPONENTS-SCORE) AND
   FIELD-POSITION > 10
THEN CALL INSIDE-RUN
```

Now, there is actually a reason for each of the factors being tested to relate to the particular decision being made here. However, there is no mention of *why* any of these factors is relevant. There is no representation of the relationship in the program, but it can nonetheless call the play at least fairly appropriately.

The COACH program calls plays in a way that takes into account the goals it is pursuing. Since its consideration of these goals is not reflected in the simple calling of the play, nominally the program's output behavior, it is important that the program be able to explain *why* it called the play it did, or why it *didn't* call some other possible play. The following is actual verbatim output from the COACH program, along with some editorial comments marked by square brackets ([ ]).

[Begin output log]

> (coach)

```
+-------------------------+
| Time: 30:00             |
| Down: 1      To go: 10  |
| Line of scrimmage: -1   |
| Home: 0    Visitors: 22 |
+-------------------------+
```

[A negative line of scrimmage means a yardline in one's own
territory, while a positive value indicates the opponent's
territory.]

>>go

I AM CALLING INSIDE-RUN
QUESTIONS? why not sweep

I DIDN'T CALL THAT PLAY BECAUSE THE SWEEP HAS A HIGHER
CHANCE OF LOSING YARDAGE THAN OTHER RUNS BECAUSE IT STARTS
SO DEEP IN THE BACKFIELD. WE DON'T WANT TO LOSE YARDAGE
NOW BECAUSE WE DON'T WANT TO GIVE THEM A SAFETY.

Play is #{Play: INSIDE-RUN}
YARDAGE is: 3

```
+-------------------------+
| Time: 29:00             |
| Down: 2      To go: 7   |
| Line of scrimmage: -4   |
| Home: 0    Visitors: 22 |
+-------------------------+
```

>>go

I AM CALLING SIDELINE-PASS
QUESTIONS? why

I CALLED THE SIDELINE-PASS BECAUSE THE SIDELINE PASS
IS GOOD FOR 4 TO 7 YARDS. WE WANT TO GET ALL THE YARDAGE
WE NEED FOR A FIRST DOWN HERE BECAUSE THE CLOCK IS WORKING
AGAINST US. WE WANT TO GET A FIRST DOWN BECAUSE WE ARE
BACKED UP AND WE NEED TO GET OUT OF HERE.

[This is not necessarily the best call. That
is due to the fact that the program has a
much smaller set of rules than it really
needs to call plays well.]

Play is #{Play: SIDELINE-PASS}
YARDAGE is: 6

```
+--------------------------+
| Time: 28:00              |
| Down: 3      To go: 1    |
| Line of scrimmage: -10   |
| Home: 0    Visitors: 22  |
+--------------------------+
```

>>go

I AM CALLING SNEAK
QUESTIONS? why not inside-run

I DIDN'T CALL THAT PLAY BECAUSE ALTHOUGH THE INSIDE RUN
IS SAFE BECAUSE LITTLE BALL HANDLING IS INVOLVED, THE
PLAY I CALLED IS SIGNIFICANTLY BETTER. WE DON'T WANT
TO LOSE THE BALL NOW BECAUSE WE WOULD BE GIVING UP A
SURE FIELD-GOAL, AT LEAST.

Play is #{Play: SNEAK}
YARDAGE is: 1

[... a play is skipped]


```
+--------------------------+
| Time: 26:00              |
| Down: 2      To go: 1    |
| Line of scrimmage: -20   |
| Home: 0    Visitors: 22  |
+--------------------------+
```

>>go

I AM CALLING SNEAK
QUESTIONS? why not long-pass

I DIDN'T CALL THAT PLAY BECAUSE THE LONG PASS HAS A FAIRLY
HIGH CHANCE OF INTERCEPTION. WE DON'T WANT TO LOSE THE
BALL NOW BECAUSE WE WOULD BE GIVING OUR OPPONENTS EXCELLENT
FIELD POSITION.

Play is #{Play: SNEAK}
YARDAGE is: 3

```
[... several plays are skipped]


        +-----------------------------+
        | Time: 22:00                 |
        | Down: 2       To go: 1      |
        | Line of scrimmage: -45      |
        | Home: 0    Visitors: 22     |
        +-----------------------------+
>>go

I AM CALLING LONG-PASS

[Note that in the previous 2nd and 1
situation the program called a safe play:
the quarterback sneak.  Now it is less
conservative due to field position.  It
chooses to pursue the touchdown goal because
the fact that a first down can probably be
gained on the next play makes obtaining a
first down on this play less crucial.]

QUESTIONS? why

I CALLED THE LONG-PASS BECAUSE THE LONG PASS IS
THE WAY TO GO FOR A T.D..  WE WANT TO GO FOR A
TOUCHDOWN NOW BECAUSE WE CAN PUNT THEM DEEP FROM
HERE SO WE CAN AFFORD TO THINK OFFENSIVELY.

[End output log]
```

The COACH program is mainly producing COHERENCY EXPLANATIONS here. It does not actually use the explanations itself at this point, although it will have to eventually deal in FAILURE EXPLANATIONS in order to achieve its goal of inventing novel plays. The program also to some extent could be said to be performing CONTRIBUTORY EXPLANATION, in that it deals in its motivations for its actions. This is mitigated, though, by the fact that the program is basing all of its reasoning on the assumption of a top-level goal of winning the football game, or at least of not losing it. It is therefore in principle incapable of simulating or understanding the reasoning of Nebraska coach Tom Osborne in this year's Orange bowl game for the national championship. In the closing seconds of the game, behind by one point, he had to chose between attempting a near-certain one-point score which would have left the game tied and won the national championship, and a fairly unlikely two-point score which would have one the game outright. He chose the two-point attempt, which failed. This was not only explainable but was

quite predictable given the fact that his team was to that point undefeated and untied, and had throughout the year been touted as possibly the best in college football history. When we have a program which can simulate and explain this kind of behavior, then we will have a program which we can legitimately say is at or beyond the level of COGNITIVE UNDERSTANDING.

Intelligence

The real intent of Artificial Intelligence, is, I claim, to find out what intelligence is all about. We tend to say that a person is intelligent to the extent that he is insightful, creative, and in general, able to relate apparently unrelated pieces of information to come up with a new way of looking at things. We tend to claim that a person is unintelligent to the extent that his behavior is thoroughly predictable with reference to what we know that he knows. Thus, when a person does things the way he was told to do them, never questioning and thus never creating new methods, we tend to see him as unintelligent.

I mention this here because I see the Explanation Game as a kind of intelligence test. We are not asking the computer to simply replicate intelligent behavior because we have no knowledge of which aspects of such behavior are more intelligent than others. Is composing a sonnet more or less an intelligent act than playing chess? There is no way to answer this objectively because it isn't the acts themselves that are at issue here, but rather the quality of those acts. Turing could not differentiate between these feats because he did not have the experience of trying to build programs to do each task. But now, as a result of years of AI research, such a question is easier to answer.

We can make a program write bad sonnets or play decent chess fairly easily. Neither of these feats seem much of a mark of intelligence. Indeed, working on either of them would not be considered AI any more, although such work might have been all right not so long ago. Today, work on computer poetry or computer chess falls within the domain of AI to the extent that it mimics the complex cognitive processes associated with the CREATIVITY inherent in both acts. Thus, if the computer poetry program started with a set of feelings and was able, by relating such feelings to its personal experiences, to create poetry, particularly poetry of some new type, we would be legitimately impressed. Similarly, if our computer chess program was capable of improving its playing ability by inventing a new strategy or employing an old one that it recalled having seen in a match it knew about, that would be an AI-type feat. The only way to distinguish this type of program from a brute-force one, though, is if the program can explain itself. Thus, if the poetry program can explain the feelings which lay behind its creation of a sonnet, or the chess program can explain how it hit upon a new strategy, we can credit that program with real understanding within its domain.

We have come to understand in AI, that it isn't the tasks themselves that are interesting. What matters is how they are done. Thus, I claim, the only way to know if our machines are intelligent is to make them do what we would expect a human to do in a similar situation. We must expect them to to be able to explain how they did it. Furthermore, those explanations should have some connection with how the task in question actually was performed. Often this is a difficult task for people to perform. We do not always know where our creative powers come from or how they were employed in any given instance. But, we can attempt to give rational explanations. We should demand no less from machines.

How The Explanation Game Works

It should be clear that there is no passing or failing of the explanation game as such. The reason for this is that the game refers to a continuum of understanding and thus it is possible to pass the test at one point and fail it at a point immediately to its right. It should also be clear that there aren't three types of explanation really. The three types I chose here conformed to the three points that I chose to label on the spectrum. (I could have chosen twelve points on the spectrum to talk about and then I would have had to come up with twelve kinds of explanation.)

The game itself is simple. It revolves around the completion of a specific task. A mental task is given to a machine on the one hand, and a person on the other, as in Turing's Imitation Game. The interviewer is asked to question the machine or person about how he came up with the behavior that he did. If the interviewer judges one subject's answers to be more insightful and explanatory than the other's, then that subject is judged to have won the Explanation Game. If that subject happens to be the machine, then we can the machine can be said to be understanding at the level of explanation that the task itself was rated. In other words the degree of passing is related to the complexity of the task. Some tasks require a depth of explanation that others do not. It is those more complex tasks that ought naturally to be pursued later in the development of AI, but, given the nature of AI, it is unlikely to happen quite that way.

The key point is that, as long as people give better explanations than machines for given tasks, then they will be naturally claimed to understand better than machines. But, when machines outstrip people in their explanatory ability, machines will be safely claimed to be better understanders, and hence more intelligent, than people in that area of knowledge. To give a simple example, machines can already out-compute humans. What they cannot do is explain the computation. Thus, while we might prefer to use a computer for our calculations, until we prefer to use a computer over a mathematician to explain the nature of the operations in mathematics, machines will not be able to pass the Explanation Game for mathematics. Here too, however,

the different points on the spectrum are operating. We may be able to enable the machine to accurately explain what it does mathematically and thus achieve the MAKING SENSE level of explanation in mathematics. But, we would have to get a machine to understand why it did what it did to come up with some new mathematical idea in order to claim the COGNITIVE UNDERSTANDING level. This is why Lenat's work (Lenat, 1977) caused such a stir. It came quite close to achieving this level of creativity/explanation ability.

The Explanation Game then, is not really a question of imitation. Our question is not so much whether a person could fool us into believing that he is a machine or whether a machine could fool us into believing that he is a person. Rather, we are interested in finding out whether anybody or anything that we talk to can be coherent in its understanding and explanation (ACCURACY above); can be creative and self-referential in its understanding and explanation (SURPRISE above); and, can be truly insightful and UNDERSTANDING (in the human sense of that word) in its understanding and explanation (EMOTION above). Machines will demonstrate such capabilities. It is just a matter of time.

References

[Colby 73]   Colby, K.M.
                Simulations of Belief Systems.
                In R. Schank and K. Colby (editors), *Computer Models of Thought and
                    Language,* . W. H. Freeman and Company, San Francisco, 1973.

[Hammond 83]   Hammond, K. J.
                Planning and Goal Interaction: The Use of Past Solutions in Present Situations.
                In *Proceedings of the National Conference on Artificial Intelligence.*
                    AAAI-83, Washington, D.C., August 1983.

[Lebowitz 80]   Lebowitz, M.
                Generalization and Memory in an Integrated Understanding System.
                PhD thesis, Yale University, October 1980.

[Lenat 76]   Lenat, D.B.
                *AM: An Artificial Intelligence Approach to Discovery in Mathematics as
                    Heuristic Search.*
                PhD thesis, Stanford University, 1976.
                Reprinted in R.Davis and D.Lenat, 1980, *Knowledge-based Systems in
                    Artificial Intelligence,* McGraw-Hill, New York.

[Restak 79]   Restak, R.M.
                The Brain: The Last Frontier.
                Doubleday and Company, Inc., Garden City, New York, 1979.

[Schank 82]   Schank, R.C.
                Dynamic memory: A theory of learning in computers and people.
                Cambridge University Press, 1982.

[Schank and Abelson 77]
                Schank, R.C. and Abelson, R.
                Scripts, Plans, Goals and Understanding.
                Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1977.

[Schank and Riesbeck 81]
                Schank, R.C. and Riesbeck, C.
                Inside Computer Understanding: Five Programs with Miniatures.
                Lawrence Erlbaum Associates, Hillsdale New Jersey, 1981.

[Turing 50]   Turing, A. M.
                Computing Machinery and Intelligence.
                Mind (59):433-460, 1950.

[Weizenbaum 76] Weizenbaum, J.
                Computer Power and Human Reason.
                W. H. Freeman and Company, Hillsdale, New Jersey, 1976.

[Winograd 72]   Winograd, T.
                 Understanding Natural Language.
                 Academic Press, New York, 1972.

DAT
FILM