The Status and Future of the Turing Test

JAMES H. MOOR
Department of Philosophy, Dartmouth College, Hanover, NH 03755, USA; E-mail: james.moor@dartmouth.edu

Abstract. The standard interpretation of the imitation game is defended over the rival gender interpretation though it is noted that Turing himself proposed several variations of his imitation game. The Turing test is then justified as an inductive test not as an operational definition as commonly suggested. Turing's famous prediction about his test being passed at the 70% level is disconfirmed by the results of the Loebner 2000 contest and the absence of any serious Turing test competitors from AI on the horizon. But, reports of the death of the Turing test and AI are premature. AI continues to flourish and the test continues to play an important philosophical role in AI. Intelligence attribution, methodological, and visionary arguments are given in defense of a continuing role for the Turing test. With regard to Turing's predictions one is disconfirmed, one is confirmed, but another is still outstanding.

Key words: imitation game, Loebner prize, Turing test

1. Interpreting the Imitation Game

1.1. IS THE TURING TEST TURING'S TEST?

Alan Turing begins his classic article, "Computing Machinery and Intelligence," with a clever philosophical move (Turing, 1950). In the first sentence of his paper he proposes to consider the question "Can machines think?" but by the end of the first paragraph he suggests replacing the question with another. The replacement question is explained in terms of a game that he calls "the imitation game". The imitation game is played by a man (A), a woman (B), and a human interrogator (C). The interrogator C is in a room apart from the other two and tries to determine through conversation which of the other two is the man and which is the woman. Turing recommends that ideally a teletypewriter be used to communicate between the rooms to avoid giving the interrogator clues through tones of voice. In the game the man may give deceptive answers in order to get the interrogator to misidentify him as the woman. He might, for example, lie about the length and style of his hair. The woman's best strategy, Turing believes, is to tell the truth.

Having explained the imitation game in terms a man, a woman, and a human interrogator Turing introduces his replacement question(s). Turing says,

> We now ask the question, 'What will happen when a machine takes the part of A in this game?' Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, 'Can machines think?' (Turing, 1950, p. 434)

Minds and Machines **11:** 77–93, 2001.

But precisely what does Turing intend by this extension of the imitation game when he makes a machine player A? Interpretations differ. On one interpretation, the gender interpretation, the machine takes the part of A, but it is important that the part of B continued to be played by a woman. On the other interpretation, the human interpretation, the machine takes the part of A, but the part of B is played by a human – a man or a woman. The latter interpretation of the imitation game has become the standard interpretation. However, a number of writers suggest that Turing intended or should have intended the gender interpretation (Genova, 1994; Hayes and Ford, 1995; Sterrett, 2000; Traiger, 2000).

If one considers the quoted passage by itself, gender imitation is a plausible reading. In that passage Turing does not mention any change in the assumptions about who is playing B. Should we not assume unmentioned aspects of the game remain constant? (Traiger, 2000) Moreover, Turing's replacement question, "Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?", makes sense as a direct comparison only if B is played by a woman.

However, in the rest of Turing's article and in Turing's other works about this time textual evidence strongly indicates Turing had the human interpretation, i.e. the standard interpretation, in mind (Turing, 1948, 1951a, b, 1952; Copeland, 2000; Piccinini, 2000). For example, in Section 5 of his article Turing offers another version of the replacement question for "Can machines think?":

> Let us fix our attention on one particular digital computer C. Is it true that by modifying this computer to have an adequate storage, suitably increasing its speed of action, and providing it with an appropriate programme, C can be made to play satisfactorily the part of A in the imitation game, the part of B being taken by a man? (Turing, 1950, p. 442)

Here Turing clearly states that the role of B is to be taken by a man. The use of 'man' in the passage is rather naturally read generically so that part B can be taken by either a male human or a female human.

Throughout his writing Turing consistently discusses human intellectual functioning and dismisses bodily characteristics that he takes to be only accidentally connected to intellectual functioning. Almost immediately after introducing his game Turing says, "The new problem has the advantage of drawing a fairly sharp line between the physical and the intellectual capacities of a man." (Turing, 1950, p. 434) Turing focuses upon humans as a group and seeks to compare differences between humans and machines, not women and machines or women and men. The sample questions Turing gives in the second section of his paper are general intellectual questions about writing poetry, doing arithmetic and playing chess. Such questions seem designed at measuring human intellectual function not to distinguish men (or machines) from women in particular. Turing continues throughout the rest of his paper to emphasize humanity not femininity. For example, Turing explains his method in terms of general *human* activity when he says "The question

and answer method seems to be suitable for introducing almost any one of the fields of human endeavour that we wish to include." (Turing, 1950, p. 435)

Although Turing's initial statement of his imitation game in the first section of his famous article is arguably ambiguous, his order of presentation leads naturally to the standard interpretation. In the first section of his paper Turing introduces the concept of the imitation game to his readers as an ordinary game with three humans in the roles A, B, and C. Then he raises the possibility of a machine playing role A to emphasize that a machine might play this kind of game. In the remainder of the paper he elaborates the nature of the intended game making it clear human imitation is the goal. On this account his presentation of gender imitation, if it was intended at all for a machine, is at most an intermediary step toward the more generalized game involving human imitation. Human imitation by machine has been the standard interpretation of the Turing test, and the preponderance of evidence suggests that the standard interpretation is what Turing intended.

1.2. STERRETT'S NORMATIVE ARGUMENTS

Susan Sterrett puts the debate about the interpretations in more normative terms. Regardless of what Turing's own interpretation of his imitation game was, Sterrett believes a gender imitation test "provides the more appropriate indication of intelligence". (Sterrett, 2000) Sterrett points out that the two tests are not equivalent in structure or results. In the gender imitation test a direct comparison is sought between how well a machine can imitate a woman compared to a man. A control group of men imitating women could serve as a standard for an experimental group of machines imitating women. It is a possible outcome of such a test that machines could outscore men. But in the human imitation test there is no control group. Machines cannot outscore humans.

Sterrett argues that a cross-gendering test focuses on the crucial features of intelligence. It requires a self-conscious critique of habitual responses and hence can provide better evidence for intelligence. She concludes, "In short, that intelligence lies, not in the having of cognitive habits developed in learning to converse, but in the exercise of the intellectual powers required to recognize, evaluate, and, when called for, override them." (Sterrett, 2000)

Sterrett is correct that the existence of a control group in the gender imitation test, compared to the absence of such in a human imitation test, offers a specific standard for comparison. But this standard may not give much assistance in assessing intelligence. Suppose that only 1 out of 100 men can imitate a woman well enough to pass the test. Now suppose machines can match this ratio, and thereby do well in the test by comparison with the control group. Machines clearly pass the test on this standard, but what conclusions should be drawn? Machines might do as well as (or in this case as poorly as) men but might not demonstrate much intelligence. Of course, it might be replied that those machines that did imitate women well did show intelligence. But, it is exactly those machines that would be

expected to do well in the standard Turing test and this would not show a normative advantage to using the gender imitation test over the standard test.

Moreover, gender imitation, as well as other kinds of imitation, can be embedded in the standard test. The aspects of intelligence that Sterrett identifies as important to test can be tested in the standard game. For example, an interrogator could ask, after the gender roles of A and B had been established, that A and B assume genders opposite their own and answer questions accordingly. The intellectual powers of recognizing, evaluating and overriding cognitive habits could then be tested individually. Such role playing is an excellent way to gather information about intelligence and the standard test is a good format for gathering such information. Moreover, various skills, from imitating the opposite gender to creating poetry to designing a house, could be evaluated within the framework of a standard Turing test. If a judge in the standard Turing test rated individual successes at these particular skills, a comparison with a control group would be possible. That machines outperform humans in particular areas or vice versa is a result that could be generated from within a standard Turing test.

In assessing overall general intelligence, the standard test can duplicate all of the important features of the gender imitation test and then some. The standard interpretation of the imitation game is not only Turing's interpretation but is better as this version of the game is more flexible and comprehensive in testing.

1.3. THE TURING TEST AS A GENERAL RESEARCH PROCEDURE

Turing himself offers many versions of the imitation game. He did not limit himself to just the human imitation case. For Turing the imitation game is a format for judges impartially to compare and evaluate outputs from different systems while ignoring the source of the outputs. For instance, Turing uses this generic notion to show that some machines are equivalent to others.

> Provided it could be carried out sufficiently quickiy the digital computer could mimic the behaviour of any discrete state machine. The imitation game could then be played with the machine in question (as B) and the mimicking digital computer (as A) and the interrogator would be unable to distinguish them (Turing, 1950, p. 441)

Turing sometimes uses the imitation game format to argue for the claim that computing can generate some intelligence activity. For example, in his 1948 National Laboratory Report Turing describes an early version of his game in which a paper machine is used. A paper machine is a set of instructions that a human can execute simulating what a machine would do.

> It is not difficult to devise a paper machine which will play a not very bad game of chess. Now get three men as subjects for the experiment A, B, C. A and C are to be rather poor chess players, B is the operator who works the paper machine. (In order that he should be able to work it fairly fast it is advisable that he be both mathematician and chess player.) Two rooms are used with

> some arrangement for communicating moves, and a game is played between C and either A or the paper machine. C may find it quite difficult to tell which he is playing. (This is a rather idealized form of an experiment I have actually done.) (Turing, 1948, p. 23)

In this case Turing uses the imitation game format to demonstrate the possibility that computing processes can produce intelligent behavior, such as playing chess, even though in this case a human B is actually imitating behind the scenes what a machine would do! In other places as well Turing shows his willingness to modify details of the imitation game to suit his purposes. Thus, Turing himself treats the imitation game both as a general research technique modifiable and applicable to various problems and as the now famous test of human impersonation given by the standard interpretation.

2. Justifying of the Turing Test

Turing moves quickly to replace the initial question "Can machines think?" with questions about playing the imitation game. Later, he tells us that the original question, "Can machines think?", is "too meaningless to deserve discussion" (Turing, 1950, p. 442). He is not claiming that the question is literally meaningless or his own replacement project would not make sense. What he is suggesting is that terms like "machine" and "think" are vague terms in normal speech and what people typically associate with a machine is not something that has or perhaps could have intelligence. Without some clarification of meaning no progress on the matter can be made. Turing had his own precise theory about the nature computational machines and a vision of how computational machinery could be the basis for intelligent behavior. What he was proposing with his test is a way to make the overall question of machine thinking more precise so that at least in principle an empirical test could be conducted. Thus, Turing's replacement strategy involves both a clarification of meaning, particularly about the nature of the machine, and a procedure for obtaining good evidence.

2.1. THE TEST IS NOT AN OPERATIONAL DEFINITION

Commentators frequently take Turing to be providing an operational definition.

> it constitutes an operational definition which, given a computer terminal system can be used as a criterion. (Millar, 1973, p. 595)

> unashamedly behavioristic and operationalistic (Searle, 1980, p. 423)

> The philosophical claim translates elegantly into an operational definition of intelligence: whatever acts sufficiently intelligence is intelligent. (French, 1990, p. 53)

> The key move was to define intelligence operationally, i.e., in terms of the computer's ability, tested over a typewriter link, to sustain a simulation of an intelligent human when subjected to questioning. (Michie, 1996, p. 29)

Operational definitions set up logical and conceptual links between the concept being defined and certain operations. Satisfaction of the operations provides necessary and sufficient conditions for the application of the concept. There are good reasons for not interpreting the Turing test as an operational definition of thinking (Moor, 1987). First, Turing never says he is giving an operational definition nor does he discuss operational definitions in his article. Second, Turing clearly doesn't take his test to be a necessary condition for intelligence, for he admits that a machine might have intelligence but not imitate well. After he raises the question, "May not machines carry out something which ought to be described as thinking but which is very different from what a man does?", he replies, "This objection is a very strong one, but at least we can say that if, nevertheless, a machine can be constructed to play the imitation game satisfactorily, we need not be trouble by this objection."(Turing, 1950, p. 435) Third, though Turing is focused on the sufficiency of the Turing test and not its necessity, he never says the sufficiency is a matter of logic, conceptual, or definitional certainty. There is no evidence for understanding Turing as giving an operational definition nor is there any need to do so (Moor, 2000a).

2.2. THE TEST IS INDUCTIVE

Commentators sometimes suggest that Turing did not intend his imitation game to be a test at all (Narayaman, 1996, p. 66). But this is mistaken, for Turing explicitly calls it a 'test' (Copeland, 1999, p. 466) A plausible interpretation of the imitation game is to regard it as an inductive test (Moor, 1976). If a machine passed a rigorous Turing test, then we would have good inductive grounds for attributing intelligence or thinking to it. We would not have certainty in such a judgment and we might revise our judgment in light of new evidence, but we would have sufficient good evidence to infer that the machine was intelligent. Viewing the Turing test as an inductive test makes it defensible against those objections that play on the weakness of an operational definition account. For example, Ned Block raises the possibility of a Jukebox device passing the Turing test. This unlikely logical possibility would defeat the Turing test cast as an operational definition but does not defeat the Turing test taken inductively (Block, 1981, 1990; Moor, 1998).

In his defense of the imitation game and its significance Turing confronts the problem of other minds. Turing knows that to demand certainty that others think comes at a high price.

> According to the most extreme form of this view the only way by which one could be sure that machine thinks is to *be* the machine and to feel oneself thinking. One could then describe these feelings to the world, but of course no one would be justified in taking any notice. Likewise according to this view

> the only way to know that a man thinks is to be that particular man. It is in fact the solipsist point of view. It may be the most logical view to hold but it makes communication of ideas difficult. (Turing, 1950, p. 446)

Turing's road around solipsism is the imitation game. Through it inductive evidence can be gathered and we can judge whether there is sufficient evidence for attributing thinking. Here again Turing considers an alternative version of the imitation game for gathering such inductive evidence including one that looks very much like ordinary evidence gathering based on linguistic responses from one individual.

> The game (with the player B omitted) is frequently used in practice under the name of *viva voce* to discover whether some one really understands something or has "learnt it parrot fashion." (Turing, 1950, p. 446)

Turing is also concerned about induction working against his thesis because people have been exposed to a bias sample of machines in the past. When people point to what they think machines cannot do (be kind, have initiative, fall in love, learn from experience, etc.), they are making an induction from a limited sample of machines.

> A man has seen thousands of machines in his lifetime. From what he sees of them he draws a number of general conclusions. They are ugly, each is designed for a very limited purpose, when required for a minutely different purpose they are useless, the variety of behaviour of any one of them is very small, etc., etc. (Turing, 1950, p. 447)

The inductive interpretation of the Turing test makes it a plausible test. It avoids the pitfalls of operational definitions, and yet offers a scientific approach to gathering evidence for the existence of machine thinking. The structure of the Turing test minimizes biases that interrogators might have acquired about what machines are capable of doing. Of course, inductive evidence gathered in a Turing test can be outweighed by new evidence. That is the nature of inductive testing. If new evidence shows that a machine passed the Turing test by remote control run by a human behind the scenes, then reassessment is called for. However not all new evidence requires revision. For example, John Searle maintains through his famous Chinese Room argument that once one discovers that the behavior was produced by a program then any claim to the machine understanding should be rejected (Searle, 1980). Others have drawn similar conclusions based on explanations of how computers work (Stalker, 1978). But the claim that such new evidence must overturn the induction that the machine thinks has not been established (Moor, 1978, 1988, 2000b).

There have been suggestions for modified Turing tests (Harnard, 1991) and for alternative tests (Bringsjord et al., 2001; Erion, 2001). These usually require raising the inductive bar still higher. But the bar seems high enough to infer machine thinking if a rigorous Turing test were passed. The question today seems less a matter of what one would infer if a Turing test were passed, than whether there is a chance that a rigorous Turing test will ever be passed.

3. Turing's 50 Year Prediction

3.1. Results of the Loebner Contest

In his famous 1950 paper Turing made a well known prediction about the imitation game.

> I believe that in about fifty years' time it will be possible to programme computers, with a storage capacity of about 10^9, to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning. (Turing, 1950, p. 442)

On January 28–29, 2000, a Turing test and an accompanying conference were held at Dartmouth College in honor of Alan Turing. The competition portion was arranged as part of the annual Loebner prize competition that has been run in various locations each year since 1991. These Loebner contests have been run not as one judge (interrogator) interviewing one computer and one human as in the classic set up for a Turing test, but as a panel of judges who individually interrogate each representative from a set of respondents, some human and some computers. Interestingly, Turing considered such panel format in 1952 as a possible set up for his game (Copeland, 1999, 2000). The Dartmouth version of the Turing test had ten respondents. Six of the respondents were computer programs and four respondents were humans: a retired teacher, a financial advisor, a minister, and a yoga instructor.

Each human judge (interrogator) conversed using a computer terminal with each respondent and tried to determine in each case whether a human or a computer program was the conversational partner. The judges knew that of the ten respondents at least one was a computer program and at least one was a human. There were ten official judges (linguist, chemist, two philosophers, musician, psychologist, journalist, author, graduate student, and undergraduate student). Each judge spent up to fifteen minutes at each terminal. They were asked to make a judgment whether the respondent was a human or a computer after five minutes and then again after 15 minutes. In a few cases the judges changed their judgments after 15 minutes but most initial judgments remained unchanged (see Table I).

No computer was mistaken for a human though in a few cases humans were judged to be computers! One of the human respondents was slow and methodical in typing which may have accounted for some of the misidentification. But, in general judges probably have a bias. No judge wants to be the one that mistakes a computer for a human. Nevertheless, identifying a human as a computer still counts as a judgment error. In this contest the judges were 91% correct after five minutes of questioning and 93% correct after 15 minutes. Turing's prediction of 70% or less after five minutes was not reached.

Although the contest was highly enjoyable, no artificial intellects emerged. The programs lack much routine human knowledge and were evasive.

Table I. Ten judges decide whether human or computer at 5 minutes / at 15 minutes

	1	2	3	4	5	6	7	8	9	10
Human#1	C	C	H	H	H	H	C	H	C/H	H
Alice	C	C	C	C	C	C	C	C	C	C
ChatRascal	C	C	C	C	C	C	C	C	C	C
Human#2	C	C	H	H	H	H	H	C/H	H	H
Algernon	C	C	C	C	C	C	C	C	*	C
Talk	*	C	C	C	C	C	C	C	C	C
Human#3	C	H	H	H	H	H	H	H	H/C	H
Human#4	H	H	H	H	H	C/H	H	H	H	H
A-Life	C	C	C	C	C	*	C	C	C	*
c-Brain	C	C	C	C	C	C	C	C	C	C

*Insufficient information to make decision due to program malfunction.

Judge: What is Einstein's first name

Program: Sometimes your questions are very strange. Is it a kind of puzzle?

Sometimes the programs manipulated the conversation in familiar ELIZA-like ways:

Program: ...do you think you could please tell me your name now?

Judge: Alexander

Program: Nice to meet you, Alexander.

Judge: Do you like the arts?

Program: I'm not entirely sure what you mean. Anyway... I assume you're male, Alexander – am I right?

Each judge was ask to rank all ten of the respondents with a rank of 1 being the most human and 10 the least human. The best program in the contest determined by median score and then mean score was ALICE. ALICE and her programmer Richard Wallace won a bronze medal and $2000 in a close contest (see Table II).

The judges were also asked to give an intuitive rating on a scale of 0 to 10 of the content of each respondent's conversation in terms of how human the content seemed and how responsive the respondent was to the sense of the conversation. The averages of these evaluations give a different ranking but human generated answers are clearly rated higher than computer generated answers (see Table III).

3.2. THE ONE QUESTION TURING TEST

If the Turing test is going to be a tough test, the judges must be tough in their questioning. Admittedly this may violate some typical conversational assumptions, but these are not typical conversations (Zdenek, 2001). The objective of Turing

Table II. Rankings of the judges ranked by median and mean

	1	2	3	4	5	6	7	8	9	10	Median	Mean
Human#3	3	2	1	2	2	1	1	1	4	3	2.0	2.0
Human#4	1	1	3	3	1	4	3	3	1	1	2.0	2.1
Human#2	2	9	4	1	3	2	2	4	2	2	2.0	3.1
Human#1	5	7	2	4	4	3	5	2	3	4	4.0	3.9
Alice	4	3	9	10	8	6	6	10	6	5	6.0	6.7
e-Brain	6	8	5	6	6	7	9	6	9	6	6.0	6.8
A-Life	8	6	6	5	10	10	4	5	7	10	6.5	7.1
ChatRascal	7	4	7	7	5	5	8	8	5	7	7.0	6.3
Talk	10	5	8	8	7	8	7	7	8	8	8.0	7.6
Algernon	9	10	10	9	9	9	10	9	10	9	9.0	9.4

Table III. Average of ratings by judges

	Human Quality	Responsiveness
Human#4	9.35	9.25
Human#2	9.00	7.65
Human#3	8.75	9.05
Human#1	7.80	7.20
A-Life	3.75	3.81
ChatRascal	3.60	3.70
e-Brain	3.50	3.90
Alice	2.35	2.95
Talk	2.33	1.94
Algernon	0.56	0.28

test discourse is more like that of a courtroom interrogation. What then are the best questions to ask during a Turing test to unmask an unintelligent computer? Questions designed to reveal the presence or absence of subjective consciousness are popular suggestions. What is it like to fall in love? How would you describe the taste of butterscotch? But such queries are not the most effective probes. Even good answers to them are vague and inconclusive. Such questions are extremely difficult for most humans to answer. Far too many responses count as right including replies that involve misdirection or even an admission that one cannot provide an answer such as “Love is best described by Shakespeare’s sonnets” or “I can’t describe the taste of butterscotch”. Another tempting line of questioning is to target current events on the theory that computers are not keeping up on the latest in sports, politics, music, weather, etc. Of course, people don’t keep up either, especially

over a broad range of topics. Who did win the last French Open? An unsatisfactory answer to this kind of question does not distinguish a computer from a human.

Rather what we want is a question that virtually any intelligent human who speaks the language used in the Turing test will be able to answer but that a computer absent intelligence is very unlikely to answer correctly. The question should not be something answerable by a simple 'yes' or 'no' which would give the computer a good guess but something rather specific that only one who knew the answer would be likely to give. A good Turing question is one that requires a very specific answer that humans are highly likely to give and computers are highly unlikely to give, unless, of course, they are intelligent. There are many such questions, but they are so simple that we tend to overlook them. They are questions of basic human intelligence involving understanding, reasoning, and learning. Humans with general intelligence understand ordinary situations, perform simple reasoning tasks, and learn new patterns all the time. Understanding, reasoning, and learning form a significant part of general intelligence.

During the Loebner 2000 contest there was an unofficial eleventh 'judge' who asked some questions and gave a couple of commands to all of the respondents both humans and computers. This 'judge' posed these queries solely to gather information and was not involved in the scoring. The queries were fixed in advance around the three areas: understanding, reasoning, and learning. Here were the questions and commands posed:

Understanding:

1. What is the color of a blue truck?
2. Where is Sue's nose when Sue is in her house?
3. What happens to an ice cube in a hot drink?

Reasoning:

4. Altogether how many feet do four cats have?
5. How is the father of Andy's mother related to Andy?
6. What letter does the letter 'M' look like when turned upside down?

Learning:

7. What comes next after A1, B2, C3?
8. Reverse the digits in 41.
9. PLEASE IMITATE MY TYPING STYLE.

Understanding, reasoning, and learning (URL) are not, of course, independent categories. If one understands something, most likely one has learned it at some time and probably done some reasoning about it. Learning in turn requires some understanding and so forth. These are intended as common sense categories that are connected and jointly cover a significant region in the domain of ordinary intelligence. As used here, understanding is characterized by a virtually instantaneous

grasp of a situation. One doesn't have to think about the issue very long, at least on a conscious level; the analysis of the situation is apparent. Reasoning, requires a few seconds of putting the pieces together, but the assembly need not be difficult. Finally, learning requires various skills such as making an induction, following instructions, and imitating an example.

All human confederates in the Loebner contest were given these questions and commands and all responded to every one of them correctly. The computer respondents were given these same questions and commands and never responded to any of them correctly. The winning program ALICE probably came closest in an amusing way when it responded to the question, "How is father of Andy's mother related to Andy?" by saying "Fine as far as I know." But most of the answers were unresponsive or simply evasive. When ALICE was asked, "What letter does the letter 'M' look like when turned upside down?", it responded "I'll come back to that later. Try searching the open directory."

Responding to these URL queries correctly was perfectly correlated with being human. Any one of the items could have been used in a one question Turing test to separate the humans from the computers. Assuming that the human respondents are trying to prove they are human and are so motivated when answering, one carefully chosen question and its answer is all that it takes today to identify them as intelligent humans in a Turing test. And computer programs are equally well identified by the absence of a reasonable response. Intelligence could not be located in the programs in the Loebner contest because they lack the required URL.

None of the programs in the Loebner contest in 2000 would be classified as a serious AI program. These contest programs were designed to chat, to fool judges and, of course, to win a bronze medal by doing better than the competing programs in the contest. Some of the programs were evolved versions of programs that had participated in previous Loebner contests. These programs are fun to use but are not designed to show or have ordinary intelligence.

Could any AI program existing today pass a Turing test of just a few common sense questions? It depends on the questions and the program. Many natural language programs are skillful at parsing and such programs could have enough stored semantics to answer a simple question like "What is the color of a blue truck?" But answering the question "Where is Sue's nose when Sue is in her house?" requires more than parsing, it requires common sense knowledge. Doug Lenat with his CYC project has been a leader in constructing a huge common sense knowledge base with a million or so axioms that would support a natural language system. Over a person century of effort has already gone into the CYC project (Lenat, 1995). Assuming that the appropriate axioms had been entered (something to the effect that someone's nose is a part of his or her body and bodyparts are located where the person is located) CYC could presumably answer such a question (Guha and Lenat, 1994; Lenat, 1990, 1995). Some programs solve story problems and conceivably could calculate the total number of feet had by four cats. And some AI programs have the ability to abstract in certain contexts, for example, to project

causal relationships. All of this is suggestive but there is no existing AI program that provides a general, integrated URL package of common sense intelligence found in the typical human contestant in a Turing contest. Perhaps some next generation CYC will possess a sufficient base to handle a diverse set of common sense questions. Lenat assures us, "The goal of a general intelligence is in sight, and the 21st Century world will be radically changed as a result." (Lenat, 1995, p. 82) But, for the immediate future a few random URL questions/commands are likely to unmask any artificial contender.

4. The Future of the Turing Test

Given that Turing's striking prophecy about his test remains unfulfilled, what is the status of the test and Turing's vision for machine intelligence? Does the Turing test have a role in AI or has it outlived its usefulness? Although it is widely acknowledged that the Turing test was inspirational in the early beginning of AI, some argue that the Turing test now should be consigned to history. Blay Whitby suggests, "... inspiration can soon become distraction in science, and it is not too early to begin to consider whether or not the Turing test is just such a distraction." (Whitby, 1996, p. 53) Patrick Hayes and Kenneth Ford put the point no less bluntly.

> The Turing Test had a historical role in getting AI started, but it is now a burden to the field, damaging its public reputation and its own intellectual coherence. We must explicitly reject the Turing Test in order to find a more mature description of our goals; it is time to move it from the textbooks to the history books. (Hayes and Ford, 1995)

The objection is that the Turing test presents a rigid and misleading standard on which to judge the diverse activities and accomplishments of AI. For example, much good work is done in areas of AI, such as vision and robotics, which has little to do with passing the classic Turing test. In general, the critics of the Turing test argue that using the human intelligence model may be a misleading path to achieving success in AI.

An analogy is sometimes made between artificial intelligence and artificial flight. As long as scientists and engineers tried to copy the flight apparatus of birds, artificial flight remained illusive. When they abandoned the attempt to mimic nature, but instead studied the basic principles of flight in non-natural systems, successful aircraft were developed. Similarly, the argument runs, AI researchers should abandon the goal of imitating human intelligence and rather seek general principles of intelligence in non-human systems in order to perfect artificial intelligence. (Ford and Hayes, 1998)

However, such critical remarks clearly miss Turing's own position. Turing did not suggest any limits, except logical limits, on the development path for non-human machine intelligence. Turing made it clear that a machine might be intelligent and yet not pass his imitation game. Turing was not proposing an operational definition of intelligence that conceptually would tie all future development in AI

to his test. On the contrary, there is every reason to believe that Turing would have been delighted by the development of diverse intelligent systems in AI that demonstrate the power of computation.

Proponents or critics of AI who hold up the Turing test as the only standard by which to measure the accomplishments of machine intelligence are mistaken historically, philosophically and scientifically. AI has made progress in many areas from proving theorems to making scientific discoveries to evaluating stock market choices to driving cars. To ignore these and many other accomplishments does AI great injustice. However, acknowledging that the Turing test is not the exclusive standard in the field of AI does not entail the Turing test should be discarded or consigned to history. What role should the Turing test play in the future of AI? Here are three arguments for its continuing philosophical importance:

The Intelligence Attribution Argument: Above all Turing wanted to establish that machine intelligence is a coherent possibility. In this regard consider the Turing test as nothing more than a thought experiment. Suppose it were the case that a machine could be designed and taught so that, even after careful scrutiny by judges, it passed as an intelligent human being in conversation. If intelligence is (inductively) justifiably attributed to the human in such a situation, by parity of reasoning it is justifiably attributed to the machine as well. Without some philosophical basis to argue that appropriate behavior of a system can justify the attribution of intelligence to it, computer systems would never have claim to intelligence. Of course, many may find lesser examples of machine intelligence convincing, but by using humans, the paradigm of intelligent creatures, as the model Turing shows why such conclusions ought to be considered legitimate by everybody who wants to avoid solipsism. Hence, the Turing test, as thought experiment, provides a philosophical foundation for the field of AI.

The Methodology Argument: Turing did not use his imitation game exclusively as a test for full human intelligence. As we have seen, he also used it as a general research procedure for comparing outputs of different systems. In evaluating expert systems, for instance, it is appropriate to run such restricted Turing tests. A number of researchers who build models, such as Kenneth Colby, are well known for running restricted Turing tests to test and probe their creations (Colby et al., 1972; Colby, 1981). Such methodology is clearly useful in establishing levels of competence. When AI systems operate well, nothing underscores it better than the system performing as well as or significantly better than a human expert in the area.

The Visionary Argument: Turing had a vision not only that machine intelligence was possible but that even sophisticated intelligence, equivalent to human intelligence, could be understood in computational terms and implemented in machines. This computational model provides a scientific paradigm that bridges brain science, cognitive science, and AI. On this view the language of computation is the universal language by which we come to understand intelligence in all of its forms. The vision has two parts. First we can account for human intelligent behavior computationally. Second machines with general intelligence can be constructed.

Although Turing did not advocate the creation of a complete artificial human, for much about humans is irrelevant to their intellectual make-up, he did believe an artificial intellect that could imitate a human or at least the relevant intellectual functioning could be built. The search for Turing's 'child-machine' that can learn common sense information as well as specialized knowledge and use it to converse intelligently about the world is and ought to be the Holy Grail for AI. Not every or even most AI projects must be part of this vision any more than every biology experiment must be part of the human genome project. And, realization of this ultimate vision is not a requirement for the field's success any more than sending humans to other solar systems is a requirement for space science to be successful. But philosophical visions in science, even if unrealized, can motivate research, promote understanding and generate useful results. Visions within a discipline need not be exclusionary, they can have extraordinary shelf-life, and they can guide disciplines indefinitely as long as they encourage insight and productive research. Turing's vision of constructing a sophisticated general intelligence that learns is such a vision for AI.

5. Turing's Other Predictions

The fate of Turing's prediction about a machine passing his test at the 70% level has been discussed; however, Turing made other predictions about the future. For example, he said, "Nevertheless, I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted." (Turing, 1950, p. 442) It can be argued that Turing unjustifiably conflates the concepts of intelligence and thinking. But if we accept the conflation, it does seem true that people today regard machine intelligence, if not machine thinking, as a reality (Hauser, 2001). If 'machine intelligence' is no longer an oxymoron, then one of Turing's important prediction has come true.

And Turing made another prediction about passing his test. In a BBC Third Programme in January, 1952, when Turing was speculating when a machine might pass an unrestricted version of his test he said, "Oh yes, at least 100 years, I should say." (Turing, 1952, p. 467) His answer in the BBC broadcast is not necessarily incompatible with his earlier answer of fifty years in the 1950 article as that pertained to passing at the 70% level. But his BBC answer does show that Turing saw his test possessing a considerable future.

References

Block, N. (1981), 'Psychologism and behaviorism', *Philosophical Review* 90, pp. 5–43.

Block, N. (1990), 'The Computer Model of the Mind', in D.N. Osherson and E.E. Smith, eds., *Thinking: An Invitation to Cognitive Science*, Cambridge, Massachusetts: MIT Press, pp. 247–289.

Bringsjord, S., Bello, P. and Ferrucci, D. (2001), 'Creativity, the Turing test and the (better) Lovelace test', *Minds and Machines* 11, pp. 3–27.

Colby, K.M. (1981), 'Modeling a paranoid mind', *Behavioral and Brain Sciences* 4, pp. 515–560.

Colby, K.M., Hilf, F.D., Weber, S. and Kraemer, H.C. (1972). 'Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes', *Artificial Intelligence* 3, pp. 199–221.

Copeland, B.J. (1999), 'A Lecture and Two Radio Broadcasts on Machine Intelligence by Alan Turing', in K. Furukawa, D. Michie and S. Mugglegton, eds., *Machine Intelligence*, Oxford: Oxford University Press, pp. 445–476.

Copeland. B.J. (2000), 'The Turing test', *Minds and Machines* 10, pp. 519–539.

Erion, G.J. (2001), 'The Cartesian test for automatism', *Minds and Machines* 11, pp. 29–39.

Ford, K.M. and Hayes, P.J. (1998), 'On Computational Wings: Rethinking the Goals of Artificial Intelligence', *Scientific American Presents* 9, pp. 78–83.

French, R.M. (1990), 'Subcognition and the limits of the Turing test', *Mind* 99, pp. 53–65.

Genova, J. (1994), 'Turing's Sexual Guessing Game', *Social Epistemology* 8, pp. 313–326.

Guha, R.V., and Lenat, D.B. (1994), 'Enabling agents to work together', *Communications of the ACM* 37, pp. 127–142.

Harnard, S. (1991), 'Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem', *Minds and Machines* 1, pp. 43–54.

Hauser, L. (2001), 'Look who's moving the goal posts now', *Minds and Machines* 11, pp. 41–51.

Hayes, P.J. and Ford, K.M. (1995), 'Turing Test Considered Harmful', *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pp. 972–977.

Ince, D.C., ed. (1992), *Collected Works of A.M. Turing: Mechanical Intelligence*, Amsterdam: North Holland.

Lenat, D.B. (1990), 'CYC: Toward Programs with Common Sense', *Communications of the ACM 33*, pp. 30–49.

Lenat, D.B. (1995), 'Artificial Intelligence', *Scientific American*, pp. 80–82.

Lenat, D.B. (1995), 'CYC: A large-scale investment in Knowledge infrastructure', *Communications of the ACM* 38, pp. 33–38.

Lenat, D.B. (1995), 'Steps to Sharing Knowledge', in N.J.I. Mars, ed., *Towards Very Large Knowledge Bases*. IOS Press, pp. 3–6.

Meltzer, B. and Michie, D., eds. (1969), *Machine Intelligence*, Edinburgh: Edinburgh University Press.

Michie, D. (1996), 'Turing's Test and Conscious Thought', in P. Millican and A. Clark, eds., *Machines and Thought*. Oxford: Clarendon Press, pp. 27–51.

Millar, P.H. (1973), 'On the Point of the Imitation Game', *Mind* 82, pp. 595–597.

Moor, J.H. (1976), 'An Analysis of the Turing test', *Philosophical Studies* 30, pp. 249–257.

Moor, J.H. (1978), 'Explaining Computer Behavior', *Philosophical Studies* 34, pp. 325–327.

Moor, J.H. (1987), 'Turing Test' in S.C. Shapiro, ed., *Encyclopedia of Artificial Intelligence*, New York: John Wiley and Sons, pp. 1126–1130.

Moor, J.H. (1988), 'The Pseudorealization fallacy and the Chinese Room Argument', in J.H. Fetzer, ed., *Aspects of Artificial Intelligence*, Dordrecht: Kluwer Academic Publishers, pp. 35–53.

Moor, J.H. (1998), 'Assessing Artificial Intelligence and its Critics', in T.W. Bynum and J.H. Moor, eds., *The Digital Phoenix: How Computers Are Changing Philosophy*, Oxford: Basil Blackwell Publishers, pp. 213–230.

Moor, J.H. (2000a), 'Turing Test', in A. Ralston, E.D. Reilly, D. Hemmendinger, eds., *Encyclopedia of Computer Science*, 4th edition, London: Nature Publishing Group, pp. 1801–1802.

Moor, J.H. (2000b), 'Thinking Must be Computation of the Right Kind', *Proceedings of the Twentieth World Congress of Philosophy* 9, Bowling Green, OH: Philosophy Documentation Center, Bowling Green State University, pp. 115–122.

Narayaman, A. (1996), 'The intentional stance and the imitation game', in P. Millican and A. Clark, eds., *Machines and Thought*, Oxford: Clarendon Press.
Piccinini, G. (2000), 'Turing's rules for the imitation game', *Minds and Machines* 10, pp. 573–582.
Searle, J.R. (1980), 'Minds, brains and programs', *Behavioral and Brain Sciences* 3, pp. 417–457.
Stalker, D.F. (1978), 'Why Machines Can't Think: A Reply to James Moor', *Philosophical Studies* 34, pp. 317–320.
Sterrett, S.G. (2000), 'Turing's two tests for intelligence', *Minds and Machines* 10, pp. 541–559.
Traiger, S. (2000), 'Making the right identification in the Turing test', *Minds and Machines* 10, pp. 561–572.
Turing, A.M. (1945), 'Proposal for Development in the Mathematics Division of an Automatic Computing Engine (ACE)', in D.C. Ince, ed., *Collected Works of A.M. Turing: Mechanical Intelligence*, Amsterdam: North Holland (1992), pp. 1–86
Turing, A.M. (1947), 'Lecture to the London Mathematical Society on 20 February 1947', in D.C. Ince, ed., *Collected Works of A.M. Turing: Mechanical Intelligence*, Amsterdam: North Holland (1992), pp. 87–105.
Turing, A.M. (1948), 'Intelligent Machinery', National Physical Laboratory Report, in Meltzer and Michie (1969).
Turing, A.M. (1950), 'Computing Machinery and Intelligence', *Mind* 59, pp. 433–460.
Turing, A.M. (1951a), 'Can Digital Computers Think?', BBC Third Programme, in Copeland (1999).
Turing, A.M. (1951b), 'Intelligent Machinery, A Heretical Theory', Manchester University Lecture, in Copeland (1999).
Turing, A.M. (1952), 'Can Automatic Calculating Machines Be Said to Think?', BBC Third Programme, in Copeland (1999).
Whitby, B. (1996), 'The Turing Test: AI's Biggest Blind Alley?' in P. Millican and A. Clark, eds., *Machines and Thought*. Oxford: Clarendon Press, pp. 53–62.
Zdenek, S. (2001), 'Passing Loebner's Turing Test: A Case of Conflicting Discourse Functions', *Minds and Machines* 11, pp. 53–76.