Turing's Two Tests for Intelligence*

SUSAN G. STERRETT
Duke University, Department of Philosophy, 201 West Duke Building, Box 90743, Durham, NC 27708, USA; E-mail: sterrett@duke.edu

Abstract. On a literal reading of 'Computing Machinery and Intelligence', Alan Turing presented not one, but two, practical tests to replace the question 'Can machines think?' He presented them as equivalent. I show here that the first test described in that much-discussed paper is in fact *not* equivalent to the second one, which has since become known as 'the Turing Test'. The two tests can yield different results; it is the first, neglected test that provides the more appropriate indication of intelligence. This is because the features of intelligence upon which it relies are resourcefulness and a critical attitude to one's habitual responses; thus the test's applicablity is not restricted to any particular species, nor does it presume any particular capacities. This is more appropriate because the question under consideration is what would count as *machine* intelligence. The first test realizes a possibility that philosophers have overlooked: a test that uses a human's linguistic performance in setting an empirical test of intelligence, but does not make behavioral similarity to that performance the criterion of intelligence. Consequently, the first test is immune to many of the philosophical criticisms on the basis of which the (so-called) 'Turing Test' has been dismissed.

Alan Turing's 1950 paper 'Computing Machinery and Intelligence' is well-known as the paper in which he proposed a practical test to replace the question 'Can machines think?'. On a literal reading of that paper, however, one finds not one, but two, such tests. The text is ambiguous regarding some details; inasmuch as the two formulations in the paper can be regarded as distinct, however, it must also be granted that Turing presented them as equivalent. My interest here is not primarily in the historical question of what Turing intended,[1] but in showing that the first test described in that much-discussed paper is in fact *not* equivalent to the second one, which has since become known as 'the Turing Test'. The two tests yield different results, and the first, neglected, one employs a better characterization of intelligence.

The first test realizes a possibility that philosophers have overlooked. It is commonly taken for granted that any test of machine intelligence that involves comparison with a human's linguistic behavior must be using a criterion of 'behavioral similarity to a paradigm case' (Churchland, 1996). But although the first, neglected, test uses a human's linguistic performance in setting an empirical test of intelligence, it does not make behavioral similarity to that performance the criterion of intelligence. Consequently, the first test does not have the features on the basis of which the test known as 'the Turing Test' has been dismissed as a failure.

Minds and Machines **10:** 541–559, 2000.

1. Claims of Equivalence of the Two Tests

In this section, I want to clearly identify the two tests I'll be comparing: what I call *The Original Imitation Game Test*, and *The Standard Turing Test*. The two tests are depicted in Figure 1, 'The Two Tests and How They Differ'. I'll also try to account for the rather common view that these two tests are equivalent. In subsequent sections, I'll show that, despite the similarity of the two tests, the first test is vastly superior; the features of our notion of intelligence it relies upon[2] include resourcefulness in dealing with unfamiliar tasks. But the appropriateness of using it as evidence of intelligence is not, as is the second test, hopelessly bogged down by considerations having to do with species-specific and culture-specific abilities, or by sensitivities to the specific skills of the interrogator.

The first test Turing proposed uses what I shall refer to as *The Original Imitation Game*. Turing used the term 'imitation game' but, as he used the term differently later, I distinguish this use of the term. In The Original Imitation Game, there are three players, each with a different goal: A is a man, B is a woman, and C is an interrogator who may be of either gender. C is located in a room apart from A and B and, during the game, knows them by the labels 'X' and 'Y'. C interviews 'X' and 'Y' and, at the end of the game, is to make one of two statements: ' "X" is A [the man] and "Y" is B [the woman]', or ' "X" is B [the woman] and "Y" is A [the man].' C's goal is to make the correct identification, B's goal is to help C make the correct identification, and A's goal is to try to fool C into making the wrong identification, i.e., to succeed in making C misidentify him as the woman. The game is set up so as not to allow C any clues to 'X' and 'Y"s identities other than the linguistic exchanges that occur within the game.[3] The first formulation Turing proposed as a substitute for 'Can machines think?' was this: '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?' (Turing, 1950, p. 434). I take Turing here to be describing the test as a sort of meta-game, of which the interrogator is unaware. This is what I shall call *the Original Imitation Game Test*.

In the subsequent discussion, Turing stated that, in turn, the question: 'Are there imaginable digital computers which would do well in the imitation game?' was equivalent to the following question: '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 program, 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 is not explicit about what the interrogator is to determine in this second version of the game, but the standard reading is that the interrogator is to determine which player is the computer and which is the man. Such a reading seems plausible enough, as the interrogator's task would be parallel (though not identical) to the task in the first version of the game, i.e., at the end of the interview, the interrogator is to state one of two things: either ' "X" is A

and "Y" is B.' or ' "X" is B and "Y" is A.', where, here, A is the computer and B is the man. The test for machine intelligence in this second version is then simply how difficult it is for an 'average' interrogator to correctly identify which is the computer and which is the man.[4] This is what I shall call the *Standard Turing Test*. Few have questioned the substitution of the Standard Turing Test for the Original Imitation Game Test. There are a few exceptions: some say Turing is ambiguous here (Moor, 1976, 1992), others that the paper is confused. (Turing's biographer Andrew Hodges (1983), and Douglas Hofstadter, a proponent of the value of the Standard Turing Test, for instance, take this approach (Hofstadter, 1981).) Some extract different tests from the paper than I have: Judith Genova (1994) focuses on the substitution of a 'species' game for a 'gender' game, and Patrick Hayes and Kenneth Ford (1995) follow her in this. Some commentators focus only on the test in the first section (Heil, 1998; Dreyfus, 1979).

Why have so many discussants accepted the slide from the first to the second formulation, though? Some do give reasons. Here is Roger Schank: 'Given that the problem is to get a computer to do as well at imitating a woman as a man, then the task is to get a computer to imitate a human as well as possible in its answers. Turing's test doesn't actually depend upon men and women being discernibly different, but on a computer's ability to be indistinguishable from a human in its responses' (Schank, 1985, p. 6). John Haugeland's treatment is similar: after giving a faithful description of the Original Imitation Game Test, he goes on to justify similarity to a human's linguistic performance as an adequate substitute, as follows. '... why would such a peculiar game be a test for general (human-like) intelligence? Actually, the bit about teletypes, fooling the interrogator, and so on, is just window dressing, to make it all properly "experimental". The crux of the test is *talk*: does the machine talk like a person?' Justin Leiber discusses the importance of impersonation in Turing's original formulation of the test, but then leaves the point aside, in speaking of passing the Turing test: '... proof positive, both psychological and legal, requires and requires no more than linguistic performance ...' (Lieber, 1991, p. 116).

These rationalizations do seize upon an important feature common to both tests: the requirement of being able to carry on a conversation with a human. For, human conversation requires – or, at least, can demand – a responsiveness and flexibility we associate with thought, and can be used to probe for knowledge of almost any subject. The sentiment is not new: Descartes, too, appealed to the ability to converse as one means of distinguishing reason from mere mechanism. It might, Descartes said, be impossible to tell a nonrational machine from an animal were the two similar in behavior and physical construction. Whereas, he argued, it *would* be possible to tell a nonrational machine from a rational creature, for '... it is not conceivable that such a machine should produce different arrangements of words so as to give an appropriately meaningful answer to whatever is said in its presence, as even the dullest of men can do' (in Descartes, 1987, p. 57).[5]

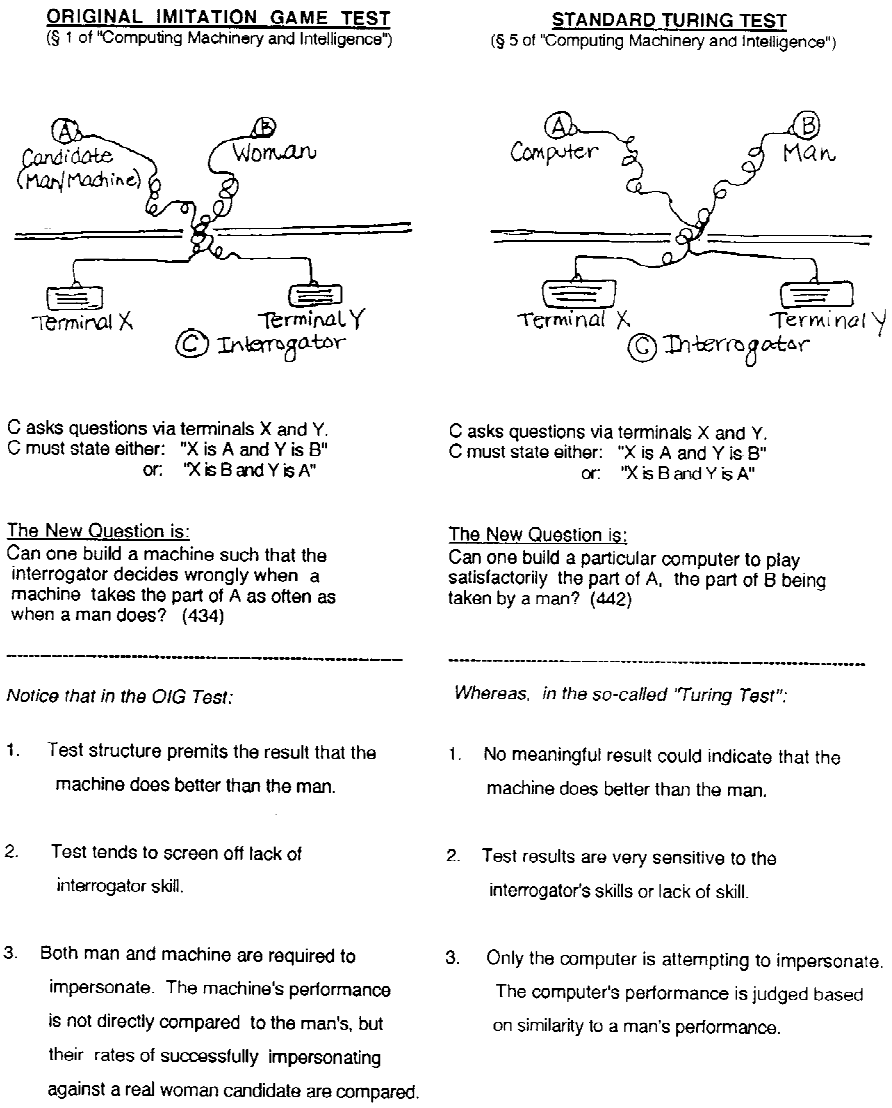

Figure 1. The two tests and how they differ.

However, while rightly identifying a common strength of the two tests, rationalizations of their equivalence overlook the distinctive difference between them. For, in spite of the fact that Turing may have thought so, too, the Standard Turing Test is not equivalent to the Original Imitation Game Test.[6]

2. Nonequivalence of the Original Imitation Game Test and the Standard Turing Test

It is not difficult to show that the two tests are not equivalent. One need only pause to consider the quantitative results each can yield. In the Original Imitation Game Test, there is nothing inherent in the game to prevent the machine from scoring higher than the man: consider, for example, the case where the man fools the interrogator into saying he is B (the woman) 1% of the time, and the machine fools the interrogator 3% of the time. In contrast, the Standard Turing Test does not admit of such a result. Recall that, in the Standard Turing Test, the interrogator's task is to identify which of the two players is the machine and which is the human. What test result would indicate that the machine had outperformed the human, given that the criterion is simply giving a performance indistinguishable from a human's? That an interrogator identified as human a machine impersonating a human (in preference to a human responding as he would in normal conversation) with a frequency greater than 50% is no measure of the machine performing a task better than the human contestants: this would mean that the interrogator has mistaken a human for a machine with a higher frequency than chance, which might well reflect more about the interrogator's quirks than about the relative capabilities of the contestants. A real-life example of how uninformative misidentifications in the Standard Turing Test can be occurred in the first Loebner restricted Turing Test: one of the interrogators mistook a human for a computer because the human exhibited what the interrogator thought a superhuman store of knowledge about Shakespeare (Schieber, 1994, p. 70).

This leads us to another difference easily exhibited by comparing quantitative results between the Original Imitation Game Test and the Standard Turing Test: the difference in the sensitivity of the test result to the interrogator's skill. The machine's fortune in passing the Standard Turing Test will go up and down with the skill level of the interrogator: if the interrogator is very poor, the percent of time the machine wins will increase; if the interrogator is very good, the percent of time that the machine wins will decrease. In contrast, the Original Imitation Game Test tends to screen off effects due to the interrogator's lack of skill. With an excellent interrogator, only extremely intelligent participants in the role of A, whether man or machine, will ever win. With a less skilled interrogator, the computer may get the interrogator to say it is the woman more often than appropriate due, say, to the interrogator's unimaginative technique; but, if C is played by the same person throughout, this will happen for the human (male) participants in the role of A as well. Since, in the Original Imitation Game Test, the machine's intelligence is measured by comparing the frequency with which it succeeds in causing the interrogator to make the wrong identification with the frequency with which a man does so, the results will not be too sensitive to the skill of the interrogator.

There are different views on the significance of the sensitivity of the test to the attitudes of the human interrogator and judge. One view is that a test's dependence

on the attitude of the human judge discredits it as a test for thinking. The construction of programs which, though very limited in the kind of responses they are capable of producing, have successfully given the illusion of carrying on one side of a conversation, have been cited to discredit the Standard Turing Test: the charge is that the test does not require enough to constitute a sufficient test for thinking. There are also charges that the dependence on similarity to a human as judged by another human requires too much: that, given a skilled and determined interrogator, only a human could pass. Yet another view is that the fact that test results in the Standard Turing Test are dependent on the human interrogator's attitude towards the candidate thinker just expresses the truism that being a thinker is best characterized as being able to be so regarded and thus represents an unavoidable aspect of any test for intelligence. Hofstadter, for instance, though a proponent of the validity of the (Standard) Turing Test, worries that 'Unless the people who play the interrogator role do so in a very sophisticated manner' the test will inspire 'a race for flashier and flashier natural-language "front-ends" with little substance behind them' (Hofstadter, 1996). My point in this paper is of significance on any of these views: *In the Original Imitation Game Test, unlike in the Standard Turing Test, scoring as a thinker does not amount to simply being taken for one by a human judge.*

3. Characterizing a Thinker – Intellectual Skill versus Cognitive Habit

In the Original Imitation Game Test, both the man and the computer are asked to impersonate. The contest between the man and the computer (as measured by the comparative frequency with which the interrogator makes the wrong identification) compares their ability to make up answers that will lead the interrogator astray. In contrast, in the Standard Turing Test, although the computer is set the task of imitating a human carrying on a conversation, the man is not called upon to imitate anything at all. Thus, what the Standard Turing Test compares is the ability of the man to converse under no pretense at all, against the ability of the computer to converse under the pretense that it is human.

Programming a computer to pass either of the two tests will involve the problem that occupied the android designers in the film *Blade Runner*:[7] giving the machine a memory, or, alternatively, the ability to fabricate responses in conversation that appear to be based on memories consistent with the normal development of a human's life. In the Standard Turing Test, all the man has to be able to do is to converse about his own memories. The analogous skill for a computer would be to converse about itself, a computer (e.g., it might say 'I became operational at the HAL plant in Urbana Illinois on January 12, 1997.') In contrast, in the Original Imitation Game Test, the man, in virtue of taking the part of player A, also needs to fabricate responses to the interrogator's questions that appear to be based on memories or knowledge consistent with having lived a woman's life. There is a great deal of intellectual dexterity and foresight involved in that task, since the

interrogator is deliberately choosing questions designed to discriminate between fabricated responses and genuine ones. And, examples of genuine responses are provided by B (the woman) at the same time that A (the man or computer) offers fabricated responses.

The skills exhibited by the man in the Standard Turing Test are just those that can be exhibited in conversation. Those skills are certainly substantial: in addition to a great deal of specific and general knowledge, conversation requires knowing what kind of responses are expected of one, and understanding the conventions that will govern the other party's understanding of one's responses. It will probably include drawing on stereotypes and presumptions that have been learned uncritically and that are used without much reflection. (Some mundane examples might be assuming that everyone eats dinner, that it can rain outdoors but not indoors, or that the conversant will understand baseball analogies such as Three strikes and you're out.') Here I mean only to be referring to features of conversation that many AI researchers and cognitive scientists have already recognized. As used in normal conversation, these skills could be called cognitive habits: they do involve cognition, but they are employed without reflecting anew on why they are appropriate each time they are employed.

However, the difficult task the man is set by the criterion used in the Original Imitation Game Test requires in addition that stereotypes get used for different purposes: rather than serving as common background drawn on in sincere efforts to communicate, they are to be used to mislead someone else to make inferences to false conclusions, which requires more reflection upon how others make inferences than is normally required in conversation. And, rather than relying on his well-developed cognitive habits in recognizing what an appropriate response would be, the man who takes the part of player A has to critically evaluate those recognitions; he has to go one step further and ask whether the response he knows to be appropriate for him is appropriate for a woman. If not, he has to suppress the response he feels is appropriate for him, and replace it with the one he determines would be appropriate for a woman, or at least one that he thinks the interrogator is likely to expect of a woman. This, I think, requires a fundamentally different ability. The point that impersonation involves intellectual abilities not necessarily exhibited by the behavior impersonated is reminiscent of Gilbert Ryle's remark about a clown's impersonations: 'The cleverness of the clown may be exhibited in tripping and tumbling. He trips and tumbles on purpose and after much rehearsal and at the golden moment and where children can see him and so as not to hurt himself ... The clown's trippings and tumblings are the workings of his mind, for they are his jokes; but the visibly similar trippings and tumblings of a clumsy man are not the workings of that man's mind' (Ryle, 1949, p. 33).

That the ability to critically edit one's recognitions of appropriate responses is required of the man by the Original Imitation Game Test, but not by the Standard Turing Test, reflects that the Original Imitation Game Test demands more of what is relevant to thinking than the Standard Turing Test does. A companion, but distinct,

point is that the Original Imitation Game Test requires less of what is not relevant to thinking, as well. This is because the task set by the Original Imitation Game Test diminishes the significance of the faithfulness of the machine's responses to a human's normal responses in determining whether the machine can think, since both man and machine are impersonating something that is neither man nor machine. This is significant because one of the challenges in designing any empirical test for intelligence (human or machine) that uses a human's performance as a benchmark will be how to screen off the peculiarities of the human or group of humans serving as the benchmark. If the test is for machine intelligence, the problem involves screening off the peculiarities of human intelligence.

These points regarding what is significant about the difference between the two tests are a bit subtle. The difference between the two tests isn't that the machine is being asked to do anything different in the Original Imitation Game Test than it is in the Standard Turing Test: the task for the machine in both tests is impersonation. However, *what counts as success* in each of the two tests is quite different. In the Original Imitation Game Test, the computer is successful if the interrogator C believes the machine's responses are sufficiently like a woman's for it to be (mis)identified as the woman with a higher frequency than C believes the performances of the man or men used in the game are sufficiently like a woman's for him to be (mis)identified as the woman. (Recall how this test is set up: in the Original Imitation Game the interrogator C is under the impression that one of the conversants is a man and one is a woman, and that the task is to say which conversant is of which gender.) Now, the machine should use whatever resources it has; it may find the best approach is to use a database of conversational exchanges and figure out what sort of features are distinctive for a woman: it might find, for instance, a species of politeness that tends to distinguish men and women, and it could fashion its linguistic response accordingly. The machine need not be successful in impersonating a woman anywhere near 50% of the time in order to pass the criterion in the Original Imitation Game. It can fail rounds based on poorly chosen responses quite often without losing at the game. What it has to do to pass the criterion of intelligence in the first test is to be sufficiently resourceful at the difficult task of impersonation to win more rounds than the man or men playing the game do. Thus, the test focuses on a notion of machine intelligence, rather than similarity to a human: that is, it focuses on the question of whether the machine is as resourceful in using *its* resources in performing a difficult task as the man is in using *his* resources in performing the same difficult task.

In contrast, in the Standard Turing Test, the machine is successful if the interrogator believes the machine's responses are sufficiently like a human's to be chosen as the human over the human. Here the kind of impersonations that yielded success in the Original Imitation Game Test may not result in success, for the interrogator is seeking a different distinction: human versus non-human. Here the problem is not that the machine doesn't have to use a self-conscious critique of its responses – of course it has to do this in both tests – but that the criterion for passing the

Standard Turing Test emphasizes things that we do not associate with intelligence, such as contingent associations between words that emphasize specifics of one's own personal experience, rather than intelligence. The machine has to fabricate these, but the man it is competing against in this test does not have to fabricate anything; thus a machine that is very resourceful (as evidenced by the first test) could lose rounds in the Standard Turing Test to a very dull man who was shown to be very unresourceful in the first test. The Original Imitation Game Test picks out the resourceful contestant; the Standard Turing Test does not.

The claim I am emphasizing here is that the criterion for a passing performance used in the first test does not penalize the machine for deficiencies that have nothing to do with machine intelligence, not the claim that producing the performances in the two tests requires fundamentally different skills of the machine. The Original Imitation Game Test requires a higher show of intelligence of the man against which to compare the machine's resourcefulness, and it does not unfairly handicap the machine by requiring undue similarity to the man's performance. Because many people have argued that a machine could never pass the Standard Turing Test, this point – that the Original Imitation Game Test does not unfairly handicap the machine – is crucial.

A helpful analogy here might be that of wanting to de-emphasize the importance of regional flavors and idioms in evaluating someone's ability to speak a second language. In a test of the ability to speak a particular language that sets up a competition between a native and non-native speaker as to who can convince an interrogator to choose him as the native speaker, an expert interrogator will be able to ferret out which is which. There will be subtle cues that can give the non-native away, no matter how well he or she has learnt the second language and become informed of various regional idioms and dialects. These cues will be due to ingrained responses that are not a matter of competence in expressing thoughts in the language. The analogous notion we are after here would be that of being recognized to be able to communicate thoughts in the language, in spite of not being able to pass under scrutiny as a native speaker of any particular region. Suppose we want to retain the approach of comparison with a native speaker. How could we use a native speaker as a paradigm case, and yet screen off the peculiarities of the region the speaker is from? Compare the following two tests as means of testing for the ability to communicate thoughts in the language: (i) competing against a native of X in being able to pass as a native of X, and (ii) competing against a native of X as to how often each of you is able to pass as a native of a different region where that language is spoken (among natives of that region). Clearly, (ii) provides better evidence of the ability to communicate thoughts in a language, as it tends to de-emphasize the importance of faithfulness to a particular regional flavor in one's expressions.

What has been missed by those not recognizing that the two tests are not equivalent is the difference in how the two tests employ the human performance in constructing a test for intelligence by which the machine's performance is to be

judged. The Original Imitation Game Test constructs a benchmark of intellectual skill by drawing out a man's ability to be aware of the genderedness of his linguistic responses in conversation. But similarity to the man's performance itself is not the standard against which the machine is compared. Without a machine in the picture at all, the man can succeed or fail at the task set. It is only successful impersonations, and then, only the fact that success has been achieved, that is the standard against which the machine is judged. That is, the measure against which the machine is judged is the frequency with which it can succeed at the same task the man is set. The man's performance does no more than normalize that measure. The successful performances of man and machine are never directly compared Oother than the fact that they are successful impersonations, similarities and dissimilarities between them are of no consequence to the test result.[8]

The significance of the use of gender in the Original Imitation Game Test is in setting a task for the man that demands that he critically reflect on his responses; in short, in setting a task that will require him to think. Gender is an especially salient and pervasive example of ingrained responses, including linguistic responses. Attempts to elicit gendered responses from us are made before we even know our own names, and continue throughout most of our lives, in interactions ranging from the most intimate to the most anonymous of interactions, from the most private to the most public of contexts. Because social interaction requires that others regard and treat someone as of a specific gender, it is well nigh impossible for someone to unilaterally ungender his interactions. Cross-gendering is not impossible, but the amount of preparation involved makes it unlikely that a player will have spent any time outside his assigned gender role. The situation is somewhat like moving in the presence of the earth's gravity; of course we are also capable of moving in 0.1 g, or 2 g as well, but we do not get opportunities to practice it. Were we suddenly put in such a situation, we would have to reflect upon the habitual components of our learned motor skills.[9] We could draw on our knowledge and observations of other bodies in motion – i.e., in this new setting, we might be more successful, even at tasks we do not normally think about, if we thought about what we were doing. Even walking might require some reflection – though still drawing on learned motor skills, we might have to reflect on how we move in order to get across the room gracefully. The Original Imitation Game Test chooses an aspect of conversation that is ubiquitous (the relevance of gender to every conversational exchange, rather than the influence of gravitational force on every physical movement), and creates a setting in which that aspect is altered so that the kind of response required is of a kind the man will not have had any practice at giving. He will not be able to rely upon his cognitive habits, and so has to figure out what response to give – in short, he has to think about it.

Thus, cross-gendering is not essential to the test; some other aspect of human life might well serve in constructing a test that requires such self-conscious critique of one's ingrained responses. The significance of the cross-gendering in Turing's Original Imitation Game Test lies in the self-conscious critique of one's ingrained

cognitive responses it requires. And, that the critique has two aspects: recognizing and suppressing an inappropriate response, and fabricating an appropriate one. It is a cliché that tests of intellectual skill differ from tests of purely mechanical skill in the novelty of the tasks set. The ability to tie a variety of knots, or to perform a variety of dives, is tested by asking the contestant to perform these tasks, and the test is not compromised if the contestant knows exactly what will be asked and practices until the task can be performed without stopping to reflect anew upon what is required. In contrast, we would think someone had missed the point of an intelligence test were the contestant given the answers to the questions beforehand, and coached to practice delivering them. The way the skills required by the Original Imitation Game Test (impersonation) differ from those required by the Standard Turing Test (conversation) is a higher-level analogue of this insight. That is, although both games involve asking questions the participant will not have knowledge of beforehand, the point is that, in the Original Imitation Game Test player A will not have had a chance to practice giving the *kind* of responses required (those that would lead the interrogator to identify him as a woman). To succeed in the task set will require drawing on knowledge of how women behave, and this cannot be a matter of relying on what I have called cognitive habit.

Nor can the task be fulfilled by simply imitating a woman's linguistic responses. Consider, for example, using a computer incorporating a connectionist net trained on a (grown) woman's linguistic responses. A little reflection on how one would select the sample linguistic responses on which to train the net shows the problem: there is no consistent way to characterize the *kind* of response here that would apply to both a woman and a machine. Clearly, using only samples from a non-game context would not be sufficient, for, in the game context, the goal of the five-minute conversation is to convince C to identify the speaker as a woman in preference to B. Thus, responses called for in the game context would not be similar to linguistic responses given in non-game conversational contexts. How about letting a woman take on the role of player A and training the net on linguistic responses she gives? The problem is that the strategy a real woman should use as player A is not the one a machine should use: A good strategy for the real woman in the role of A would be to look for opportunities to turn to a topic that exhibits her knowledge of things only a woman would know. Such a strategy will get the machine in trouble, as it is unlikely that being trained on linguistic responses alone will be a good basis on which to deal with topics turned to for the sole purpose that they are the sort of thing only a woman would know. A good strategy for an impersonator is just the opposite: to steer the conversation away from lines of questioning that might lead to a topic that would expose his ignorance of things only a woman would know. The general point here is about impersonation in contrast to imitation, not about approaches using connectionist nets (as there are other, undeveloped approaches, such as attempting to encompass a birth-to-adulthood process): it is that impersonation in contexts where one's identity is in question is not the same as imitation in normal contexts. Similar remarks apply to using the suggested approach for player A in the Standard

Turing Test (i.e., preparing the machine to pass as a human by equipping it with a net trained on a man's linguistic responses.) The additional step taken by many contemporary authors in regarding any test for linguistic competence a suitable substitute shows a disregard for the significance of the game context. I discuss other games that demand self-conscious critique of one's responses in Section 5.

4. Critical Thought and Supercritical Minds

The point of the previous section was that the self-conscious critique of one's ingrained responses required in the Original Imitation Game Test is crucial to its value as a test of intelligence. Impersonation is the context in which such self-conscious critique is most clearly exhibited, but it is also true that in other contexts such self-conscious critique marks the difference between a response requiring thought and a response that, though entirely appropriate, is habitual. It is not that habitual responses do not involve any cognitive component at all, but focusing on a human's ability to evaluate and fabricate responses is an attempt to tease out the intellectual component of linguistic responses. The purpose of the cross-gendering is to de-emphasize training, and emphasize thinking.

Now, with the point that the Original Imitation Game Test de-emphasizes training in mind, consider R.M. French's criticism of what I have called the Standard Turing Test. His view is that '... the [Standard] Turing Test provides a guarantee not of intelligence but of culturally-oriented human intelligence' (French, 1990, p. 54). French composes clever test questions to be asked in a Standard Turing Test that would distinguish human from non-human, such as 'Rate *pens* as *weapons*' and 'Rate *jackets* as *blankets*.' (p. 61). He explains why a computer would not have a chance: such questions '... [probe] the associative concept (and sub-concept) networks of the two candidates. These networks are the product of a lifetime of interaction with the world which necessarily involves human sense organs, their location on the body, their sensitivity to various stimuli, etc.' (p. 62). His point in making this remark is that there is something he calls a 'subcognitive substrate' that can be made to exhibit itself under interrogation, and that what is wrong with the (Standard) Turing Test is that it would require actually having lived a human's life, with human sensory capabilities, in the same culture as the human against whom the participant is competing, to pass it, i.e., to give linguistic responses indistinguishable from those a human gives.

Such criticisms do not have the same force against the Original Imitation Game Test, in which both the man and the computer fabricate responses. French's criticisms turn on the fact that the (Standard) Turing Test is based on comparing how well a computer would do against a human in behaving like a human. The advantage of Turing's first formulation of the test is that it provides a context in which the computer and the man are put on a more equal footing: both the computer and the man will need to critically evaluate their responses, and fabricate appropriate ones that are based on vicarious experiences of womanhood. That 'some subcognitive

substrate is necessary to intelligence' suggests that an intelligence test should be constructed such that, although the contestant is called upon to make use of such a substrate in a way that exhibits the kind of resourcefulness humans display when making use of their distinctively human subcognitive substrate in conversing, the particular substrate the candidate has will not be relevant to doing well on the test. As we have seen, the Original Imitation Game Test provides just such a screening off of the particularity of the substrate. The Original Imitation Game Test takes a step towards meeting French's challenge that no attempt to 'fix' the Standard Turing Test could address the problem that it could only be a test of human intelligence, for it retains the insight that conversation is the context in which the flexibility of response we associate with thinking is best displayed, but changes the task (from conversing to impersonating) so that faithfulness to a human's natural responses is de-emphasized. I do not claim that the man has no advantage in some respects – obviously a man will have more in common with a woman than a computer would – but I do claim that the Original Imitation Game Test takes the approach one should use to de-emphasize the human's advantage in participating in a test intended to measure general, and not merely human, intelligence. Impersonation of something that is neither man nor machine is just the task to set. What to choose that would be especially thought-provoking for a man to impersonate? I think setting the task of cross-gendering one's responses is a stroke of genius.

Odd as it may sound to those who know him only as the proponent of the so-called 'Turing Test' of intelligence, Turing actually offered reasons that human intelligence should not be considered the standard of intelligence, as well. In the 1950 paper, he counters Lady Lovelace's objection that 'the machine can only do what we tell it to do' with the simile of a nuclear reaction: 'the disturbance caused by ... an incoming neutron will very likely go on and on increasing ... Is there a corresponding phenomenon for minds, and is there one for machines?' But, surprisingly, he does not identify the ability to go critical with being a mind, or, even, a human mind; he answers instead: 'There does seem to be [a corresponding phenomenon] for the human mind. The majority of them seem to be 'subcritical' (Turing, 1950, p. 454). Not only did he judge that most humans did not exhibit whatever feature of minds is analogous to supercriticality, but he speculated that some exhibitions of machine intelligence would be superhuman. In an essay published posthumously, he impishly states that intellectuals would probably be mistaken to fear being put out of a job by intelligent machines, because 'There would be plenty to do in trying, say, to keep one's intelligence up to the standard set by the machines, for it seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers.' One reason to expect this is that the machines 'would be able to converse with each other to sharpen their wits' (Turing, 1996, p. 259).

My point does not turn on the historical question of whether or not Turing held the view that human intelligence is not the standard by which machine intelligence should be measured, though. The conceptual point is that the Original Imitation

Game Test does not use human intelligence as a standard of intelligence. To be sure, it uses comparison with a man's performance, but, as I have argued, it constructs a setting in which the man really has to struggle to produce the responses required to succeed at the task set, in which he is forced to think about something he may not otherwise ever choose to think about.[10]

You might say that the Original Imitation Game Test extracts a 'supercritical' kind of thought from humans, and then uses it to construct a measure by which the machine's capabilities can be measured. That the task of cross-gendering one's responses is so demanding might bring into question its suitability as a general test for thinking, i.e., should not a good test for thinking set a task that all and only thinkers can accomplish? Actually, that is not what either of the practical tests Turing suggested are meant to do. Rather, as numerous people have pointed out, the purpose of replacing the question concerning the possibility of machine intelligence with a practical test was to identify the sort of evidence that most reasonable people would consider evidence of intelligence. Here one should also keep in mind that success in the test is not a matter of succeeding or failing one interview, but of succeeding with the same frequency as a man or a suitably chosen sample of men would. The measure works like comparisons of batting averages; though most baseball players would not get a hit given just one pitch, and some may not ever get a hit, batting averages are still useful measures. In our case, the measure is used qualitatively: if the candidate attains an average equal to or exceeding that of a certain kind of participant under demanding conditions, the test result is taken to provide good evidence that the candidate is intelligent. This need not entail the claim that the measure would be useful in ordering people with respect to intelligence, nor that falling below the average of a certain kind of participant should be considered good evidence that the participant is *not* intelligent.

It is worth noting, however, that, were the test used as a sort of scale, not even the best performance of which a human is capable would be the ultimate standard of intelligence; for, recall that, as I have shown in Section 2, in The Original Imitation Game Test, the structure of the test allows for a result in which a machine can outperform a man. The Original Imitation Game Test is remarkable for testing the ability to think without resorting to mere anthropomorphism. Of course there is anthropomorphism involved in the human judge's acceptance of the machine's successful impersonations of a woman. But, the test is not simply a matter of a human judging similarity of a machine's behavior to a human's behavior. It tests instead for an awareness of what's involved in making inferences, especially about social conventions and what are often called social constructions (e.g., gender), by setting a test of resourcefulness in using that awareness.

5. Differentiating Human Performances

I opened this paper with the claim that the first of Turing's two tests was based on more general, and hence more appropriate, features of intelligence. It would

be discouraging to think that the ability to deceive is the ultimate intellectual achievement, or, even, somehow essential to thinking.[11] But no such conclusion is forced. Rather, the feature of deceiving the interrogator in the Original Imitation Game combines two separate features, neither of which is particularly related to deception: (i) knowing how to use the knowledge that someone else knows how to draw conclusions, and (ii) the ability to edit one's own responses.

Knowing how to use the knowledge that someone else knows how to draw conclusions is just as useful in communicating effectively – i.e., to lead one's conversant to true conclusions, rather than to mislead him or her to false ones. Of course we use such knowledge, implicitly and unreflectively, in normal conversation; we are usually only forced to reflect upon *how* we use such knowledge, however, in special situations such as educating or persuading, where the one being educated has not yet been initiated into the field with which we are familiar, or the person being persuaded does not yet share our view. There do exist parlour games that test for one's resourcefulness in employing this ability to communicate with, rather than mislead, someone. The game of charades is an example; here, the need for resourcefulness arises from the almost opposite constraint that one must use gestures rather than words to communicate. Other examples are *Password* and *Pictionary*, both of which require a player to communicate, but constrain the player from communicating a given idea or phrase in the manner he or she would normally use. The player is required to be resourceful in using what he or she knows about how others make associations and draw inferences, including exploiting icons, preconceptions, stereotypes, and prejudices as well as particular facts and specialized knowledge.

The ability to edit one's responses is just as usefully employed in behaving well as it is in deception, but, similarly, tends to be required only in special cases, i.e., ones that are sufficiently novel such that we have not had sufficient experience to have developed habitual responses. We have to be able to see when the situation calls for overriding a habitual response. Although my point does not turn on what Turing actually thought, it is noteworthy that he wrote that the machine should be constructed so that its behavior was not completely determined by its experience. In explaining a suggestion for employing random numbers, he says: This would result in behavior of the machine not being by any means completely determined by the experiences to which it was subjected. ... (Turing, 1996). The task of impersonating someone who has occupied a role in which we have had no experience at all is different from the task of imitating someone accurately as a result of practice.

I find the first proposal inspired, whatever Turing's attitude towards it may have been. Besides the formal points made above that the Original Imitation Game Test does not have the weaknesses for which the Standard Turing Test has been discredited as a test of machine intelligence, there is the additional point that it has some virtues when used to evaluate human intelligence as well. Not all human linguistic performances *should* count as evidence of intelligence; there are even cases where we would want to say: 'I must not have been thinking when I said that.' It is all

to the credit of a test for machine intelligence that it does not regard dull human performances as providing good evidence of intelligence, and that it requires one to reflect upon ingrained responses. The Original Imitation Game Test has these virtues; the Standard Turing Test does not.

It is pretty generally agreed among philosophers that the (so-called) 'Turing Test' fails as a practical test. Blay Whitby recently delivered something like a respectful eulogy for the test (Whitby, 1996), giving his opinion as to what it ought to be remembered for. Although I have tried to show that the difference between the two formulations of the test given in Turing's 1950 paper is relevant to the applicablity of the most common criticisms, I do not disagree with Whitby's statement that 'the last thing needed by AI *qua* science is an operational definition of intelligence involving some sort of comparison with human beings', or, even, his admonition that 'AI *qua* engineering should not be distracted into direct copying of human performance and methods'. What I challenge is the common presumption that any test employing a human's performance could be employing no other notion of intelligence than specifically human intelligence. I think this general point worthwhile even if the Standard Turing Test had never been discussed.

For, the importance of the first formulation lies in the characterization of intelligence it yields. If we reflect on how the Original Imitation Game Test manages to succeed as an empirical, behavior-based test that employs comparison with a human's linguistic performance in constructing a criterion for evaluation, yet does not make mere indistinguishability from a human's linguistic performance the criterion, we see it is because it takes a longer view of intelligence then linguistic competence. 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.

Notes

*I wish to thank John McDowell and Clark Glymour for encouragement and helpful discussion on early versions of this manuscript. Thanks also to Thomas Stuart Richardson, Teddy Seidenfeld, Dirk Schlimm, Peter Spirtes, Gualtiero Piccinini, Patrick Giagnocavo, Martha Pollack, Jim Moor and Jerry Massey, and to audiences at: Occidental College, Pitzer College, "The Future of the Turing Test" conference held at Dartmouth College on January 29, 2000; the 1999 Computing and Philosophy Conference (CAP '99) held at Carnegie-Mellon University August 1999, and the Theoretical Cognition Group at the University of Pittsburgh. Thanks also for helpful comments from anonymous referees.

[1]Gualitero Piccinini (2000, this volume) argues that Turing had only one test in mind. His reasoning for this grants a minor (chauvinistic) slip on Turing's part in the 1950 paper and appeals to other documents by and about Turing, including an interview with Turing held after he had written the 1950 paper I discuss here. On the other hand, Saul Traiger (2000, this volume) provides considerations against reading the 1950 paper as proposing what has become the standard version of the 'Turing Test'.

[2]I do not mean to imply that the purpose of the test is to define a notion of intelligence, or, even, to define a measure of general intelligence. I think James Moor is correct in arguing that Turing was not providing an operational definition of intelligence; he says that the value of the game

lies in its potential for providing "good inductive evidence for the hypothesis that machines think" (Moor, 1976, p. 249). I think Turing was providing an example of the kind of evidence on which it might be natural to speak of a computer as being intelligent. My comments throughout this paper regarding "the notion of intelligence" involved in certain tests should be understood in light of this distinction between providing a definition of intelligence and defining a test that provides evidence of intelligence. Obviously, the latter involves having an idea of the features that characterize intelligent behavior, or at least some examples of intelligent behavior, but it does not require that one can define intelligence, that there is a single measure of intelligence, or that the intelligence of different intelligent beings can always be compared. The fact that the answer given to the specific question of whether machines could exhibit intelligence was in the form of defining a practical test shows, I think, some respect for the problems that would be involved were one to attempt to define intelligence or thinking.

[3]Turing suggested using a teleprinter for the communication. Although, literacy is not required of a player: Turing also specified that an intermediary may communicate the answers from A and B to C.

[4]Turing made a prediction in terms of the per cent chance that an average interrogator would have of making the right identification after five minutes of questioning.

[5]In their paper "Descartes's Tests for (Animal) Mind" Gerald Massey and Deborah Boyle (1999) examine Descartes's Action Test as well as his Language Test. They argue that both are tests for mind, but that the tests differ in that the Action Test is a test for volitions, whereas the Language Test is a test for perceptions. I recommend this interesting paper to the reader who wishes to understand Descartes's views on the issue. Recall that, since, for Descartes, animals are sophisticated machines, Descartes's considerations on the question of what would consitute evidence of animal mind are particularly germane to the contemporary question of what would constitute evidence of machine intelligence.

[6]Although commentators have not generally challenged Turing's claim that the tests are equivalent, many have found the test descriptions wanting. A. Hodges regards Turing's first mention of the imitation game an uncharacteristic lapse of lucidity, in *Alan Turing: The Enigma* (New York: Simon & Schuster, 1983, p. 415). D.R. Hofstadter misunderstands the first imitation game described as a test for femininity and criticizes it as such, in 'A Coffeehouse Conversation, *Scientific American* (May, 1981, 15–36). In his encyclopedia entry, Moor (1992) does clearly describe the standard Turing Test as a variation of the imitation game, and discusses the ambiguities in Turing's 1950 exposition of it. Ford and Hayes (1995) follow Genova (1994) in regarding the difference between the standard formulation and the imitation game discussed in the opening of Turing's paper as one of determining species rather than gender, but regard the tests as having the same basic structure. (Whereas, the two tests I identify have a different structure, as evidenced by the fact that the man and machine's abilities to impersonate are compared in the first game, but not in the second.) Perhaps because of their focus on the first part of Turing's 1950 paper, Ford and Hayes (1995) take Turing to have presented only the first version. ('Turing is usually understood to mean that the game should be played with the question of gender (e.g., being female) replaced by the question of species (e.g., being human)... We will call this the species test. ... However, Turing does not mention any change to the rules of the imitation game ..." (p. 972)). Heil (1998, pp. 110–111) likewise mentions only the first formulation of the test, given in Section 1 of Turing's 1950 paper.

[7]The screenplay for the film *Blade Runner* is based on P.K. Dick's (1982) novel.

[8]This last point can perhaps be made clearer by considering the following objection: 'Why,' a resourceful reader might ask, 'could not impersonation be incorporated into the Standard Turing Test setup as follows: the interrogator could preface a set of questions with "Suppose you were a woman. How would you answer ..." Then, the interrogator would identify whichever player gave the most convincing performance as the man, and the other as a machine, thus effectively redefining 'success' as giving, in the opinion of an interrogator, the better of two impersonations. In such a version of the Standard Turing Test, the interrogator would be comparing the man's attempt to impersonate a woman with the machine's attempt to impersonate a woman. However, the context differs signific-

antly from that of the Original Imitation Game Test. The interrogator in the Original Imitation Game Test is never directly comparing two fakes against each other and picking out which he judges to be the better counterfeit; he is trying to make a correct determination of which interviewee is a woman and which interviewee is an impersonation of a woman. Asking the interrogator in a Standard Turing Test setup, in which it is known that the pair either does not or may not contain a woman, to merely pretend to focus on the task of making a gender distinction is not likely to be any more effective than asking the interrogator to merely pretend there is a screen between him and his interviewees, in lieu of actually incorporating one into the game.

Thomas Stuart Richardson has convinced me that it is not only best, but probably necessary, to specify that the interrogator in the Original Imitation Game Test be under the impression that each X, Y pair of which he is to judge either ' "X" is a man and "Y" is a woman' or ' "X" is a woman and "Y" is a man' actually does consist of exactly one man and exactly one woman.

[9]Donald Michie calls these "subcognitive skills" and remarks: 'Only when a skilled response is blocked by some obstacle is it necessary to "go back to first principles" and reason things out step by step', in Michie (1993). He recognizes that the (Standard) Turing Test requires that the machine impersonate, rather than give the answer one would give if not playing the game, inasmuch as he notes Turing's remark that 'The machine ... would not attempt to give the right answers to the arithmetic problems. It would deliberately introduce mistakes in a manner calculated to confuse the interrogator.' Michie calls this a 'playing dumb' tactic, though, and dismisses it with '... surely one should judge a test as blemished if it obliges candidates to demonstrate intelligence by concealing it!' This misses the point that such a response should not be characterized as 'playing dumb', but as impersonation. For, in this case, the machine does not make arbitrary mistakes, nor perform calculations as a human would; what the machine does, and does without error, is 'introduce mistakes in a manner calculated to confuse the interrogator.' That impersonation displays recognition of habits, or rules, without articulating them seems to me particularly germane to Michie's discussion of the articulation of rules.

[10]Both Daniel Dennett (1998, p. 23) and Justin Leiber (1991, p. 110), have remarked on the ingenuity the first game requires, though they draw different conclusions as to what success in the game would show.

[11]Some have suggested that the fact that the task set the candidate involves deception is significant. One example is Richard Wallace's 'The Lying Game' (1997).

References

Churchland, P.A. (1996), 'Learning and Conceptual Change: The View from the Neurons', in A. Clark and P.J.R. Millican, eds., *Connectionism, Concepts and Folk Psychology: The Legacy of Alan Turing*, Vol. 2, Oxford: Clarendon Press.

Dennett, D.C. (1998), 'Can Machines Think?' in *Brainchildren*, Cambridge, MA: MIT Press.

Descartes, R. (1987), *Discourse on Method*, Cottingham, J. (Trans.). Cambridge: Cambridge University Press.

Dick, P.K. (1982), *Do Androids Dream of Electric Sheep?*, New York: Ballantine Books.

Dreyfus, H.L. (1979) *What Computers Can't Do*, Revised Edition. New York: Harper Colophon Books.

French, R.M. (1990). 'Subcognition and the Limits of the Turing Test', *Mind* 99.

Genova, J. (1994), 'Turing's Sexual Guessing Game', *Social Epistemology* 8(4), pp. 313–326.

Gunderson, K. (1964), 'Descartes, LaMettrie, Language, and Machines', *Philosophy* 39, pp. 193–222.

Haugeland, J. (1985), *Artificial Intelligence: The Very Idea*, Cambridge: MIT Press.

Hayes, P. and Ford, K. (1995), 'Turing Test Considered Harmful', *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCA195-1)*. Montreal, Quebec, Canada. pp. 972–997.

Heil, J. (1998), *Philosophy of Mind: A Contemporary Introduction*. London and New York: Routledge.

Hodges, A. (1983), *Alan Turing: The Enigma*, New York: Simon and Schuster.

Hofstadter, D.R. (1981), 'A Coffeehouse Conversation', *Scientific Americans*, May 1981, pp. 15–36.

Hofstadter, D.R. (1985), *Metamagical Themas*, New York: Basic Books.

Hofstadter, D.R. (1996), 'Analogy-Making, Fluid Concepts, and Brain Mechanisms', *Connectionism, Concepts, and Folk Psychology: The Legacy of Alan Turing. Vol. II*, Oxford: Clarendon Press, pp. 195–247.

Leiber, J. (1991), *An Invitation to Cognitive Science*, Cambridge, MA: Basil Blackwell.

Massey, G.J. and Boyle, D.A. (1999), 'Descartes's Tests for (Animal) Mind' (forthcoming, *Philosophical Topics* 27, Special Issue on Zoological Philosophy and Philosophical Ethology).

Michie, D. (1993), 'Turing's Test and Conscious Thought', *Artificial Intelligence* 60, pp. 1–22.

Moor, J.H. (1992), 'Turing Test', *Encyclopedia of Artificial Intelligence*, 2nd Edition, New York: John Wiley & Sons, pp. 1625–1627.

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

Piccinini, G. (2000), 'Turing's Rules for the Imitation Game', *Minds and Machines* 10, pp. 573–582.

Ryle, G. (1949), *The Concent of Mind*, Chicago: University of Chicago Press.

Schank, R. (1984), *The Cognitive Computer*, Reading, MA: Addison-Wesley.

Shieber, S.M. (1994), 'Lessons From a Restricted Turing Test', *Communications of the ACM*; 37(6).

Traiger, S. (2000), 'Making the Right Identification', *Minds and Machines* (this volume).

Turing, A.M. (1950), 'Computing Machinery and Intelligence', *Mind*, 59, pp. 433–460.

Turing, A.M. (1996), 'Intelligent Machinery, A Heretical Theory', *Philosophia Mathematica*, (3)4, pp. 256–260.

Wallace, R. (1997), 'The Lying Game', *Wired*, Vol. 5, No. 8, August 1997.

Whitby, B. (1996), 'The Turing Test: Al's Biggest Blind Alley?' in P.J.R. Milllican and A. Clark, eds., *Machines and Thought: The Legacy of Alan Turing. Vol. 1*, Oxford: Clarendon Press.