International Journal of Modern Physics C
Vol. 15, No. 8 (2004) 1041–1047
© World Scientific Publishing Company

TESTING THE TURING TEST — DO MEN PASS IT?

RUTH ADAM

Cognitive Science Program
Hebrew University, Jerusalem, Israel

URI HERSHBERG*

Interdisciplinary Center for Neural Computation
Jerusalem, Israel
uriher@cc.huji.ac.il

YAACOV SCHUL

Department of Psychology
Hebrew University, Jerusalem, Israel

SORIN SOLOMON

Racah Institute for Physics
Hebrew University, Jerusalem, Israel

Received 23 December 2003
Revised 30 March 2004

We are fascinated by the idea of giving life to the inanimate. The fields of Artificial Life and Artificial Intelligence (AI) attempt to use a scientific approach to pursue this desire. The first steps on this approach hark back to Turing and his suggestion of an imitation game as an alternative answer to the question "*can machines think*?".[1] To test his hypothesis, Turing formulated the Turing test[1] to detect human behavior in computers. But how do humans pass such a test? What would you say if you would learn that they do not pass it well? What would it mean for our understanding of human behavior? What would it mean for our design of tests of the success of artificial life? We report below an experiment in which men consistently failed the Turing test.

Keywords: Turing test; artificial intelligence; artificial life; sentience; imitation; deception; thinking machines.

1. Introduction — What is the Turing Test?

The Turing test has stood for the last half-century as the standard test to be passed by a machine in order to be considered sentient (for two monographs which appeared recently on the subject, see Refs. 2 and 3).

*Corresponding author.

At its base, the Turing Test (TT) is an articulation of mankind's eon-long fascination with causing the inanimate to imitate life. From the Golem of Prague through Frankenstein, from Pygmalion and Galitea to Deep Blue, humans have been fascinated by the idea of bestowing life on coarse matter. TT has been celebrated in thousands of writings from scientific (psychology, computer science, philosophy, cognition, etc.) through popular science to science-fiction as the threshold separating the mechanical from the cognizant. Designing a machine that can pass the test is one of the most enduring challenges for the scientific community and the subject of a long chain of prestigious competitions (the latest of which is the Turing tournament currently organized at CALTECH[4,5]).

The interest in the Turing Test is not decaying with time and it is today as alive as ever. To quote one of the many comprehensive reviews appeared with the occasion of the half century anniversary of the TT[6]:

> "*The Turing Test will remain important, not only as a landmark in the history of the development of intelligent machines, but also with real relevance to future generations*".
>
> Admittedly some[7] think that:
>
> "*The AI establishment has for more than a decade put more energy into explaining why the Turing test is irrelevant than it has into passing it*".
>
> Yet others[8] maintain that:
>
> "*The TT simply represents what it is that AI must endeavor eventually to accomplish scientifically*".

In this paper we (try to) steer away from the very passionate and colorful confrontations decorating the TT and adopt a matter-of-fact constructive and hopefully — instructive approach. We have taken our previous experience with giving operational definitions to abstract concepts, such as creativity, perception and representation,[9–11] and tried to apply it to the TT. In these studies we empirically researched the actual behavior in humans before attempting to arrange them in a theoretical framework or comparing them to an artificial constructed system. In the following pages we show that such an attitude leads to interesting conclusions for the TT as well. What Turing touched upon so well with his test is that in human eyes the answer to "Can machines think?" would best be answered by studying the question "Can machines behave human?"

Turing suggested that machines should be tested for their ability to deceive humans into thinking that they (the machines) are human. He claimed that machines that pass this test should be considered sentient. But how do humans pass such a test? What would you say if you would learn that in fact humans do not pass well this test? Would you doubt human sentience? Or perhaps the humans' ability to administrate the humanity test? We report below an experiment in which man consistently failed the Turing test.

But first, what is the Turing test? Turing wrote[1]:

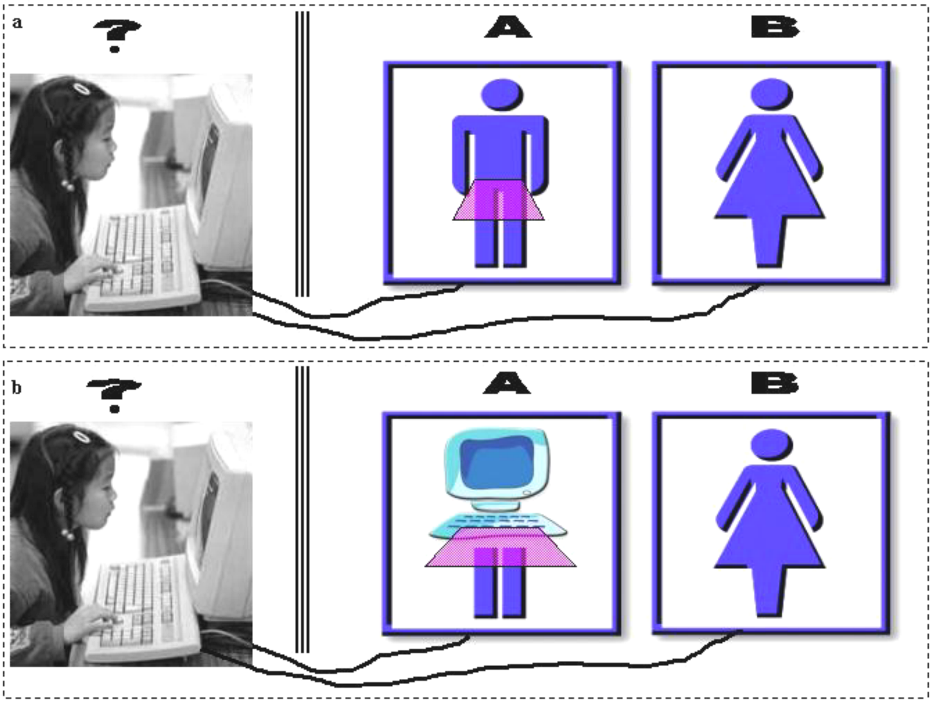


Fig. 1. Turing's imitation game: (a) the gender imitation game: the man (A) has to impersonate a woman. The woman (B) answers truthfully. The interrogator must attempt to identify correctly who is not a woman. (b) The Turing test: a machine gender test: the computer replaces the man (A) and must fool the interrogator.

> "*I propose to consider the question* '*Can machines think*?' ... *I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words. The new form of the problem can be described in terms of a game which we call the* '*imitation game*' ".

Turing's imitation game (Fig. 1(a)) has three participants: A man (A), a woman (B) and an interrogator. The interrogator asks the man and the woman identical questions aiming to identify their gender, namely, who is the woman and who is the man. But this is not so simple because the man (A) aims to fool the interrogator, while the woman (B) is truthful. In other words the man pretends that he is a woman. The interrogator is only allowed to ask them questions through a neutral interface (words written on a screen). She can ask as many questions as she wants, as often as she wants, and about any subject. Then, to construct his test, Turing continues by substituting the man (A) with a machine[1] (Fig. 1(b)):

> "*We now ask the question* '*What will happen when a machine takes the part of A* [*i.e., the man*] *in this game*?' *Will the interrogator decide wrongly*

> *as often when the game is played like this as [s]he does when the game is played between a man and a woman? These questions replace our original, 'Can machines think?'* ".

Since this suggestion was made there have been many attempts to build a computer program that would win the imitation game and trick a human interrogator. There are famous examples of programs like Eliza[12] and Racter[13] which penetrated the general culture.[2,3] However during the 53 years since Turing's article no one has ever questioned the validity of Turing's basic assumption, namely, that humans do pass the Turing test. Can a man fool the interrogator while masquerading as a woman? It is interesting why the first part of the Turing experiment — testing a man in the role of A — was never performed until now.[2,3,14] Speculatively, there are two possible explanations for this.

First: Turing and his followers, took for granted that men would perform the A role successfully. Second: people treated the first part (the human performance) as merely a benchmark for the machine performance. Since no machine came close to fulfilling the A role successfully, it appeared pointless to perform at all the first part of the test. However, the human test does become relevant once we discover that men perform below Turing's expectations for "thinking" machines.

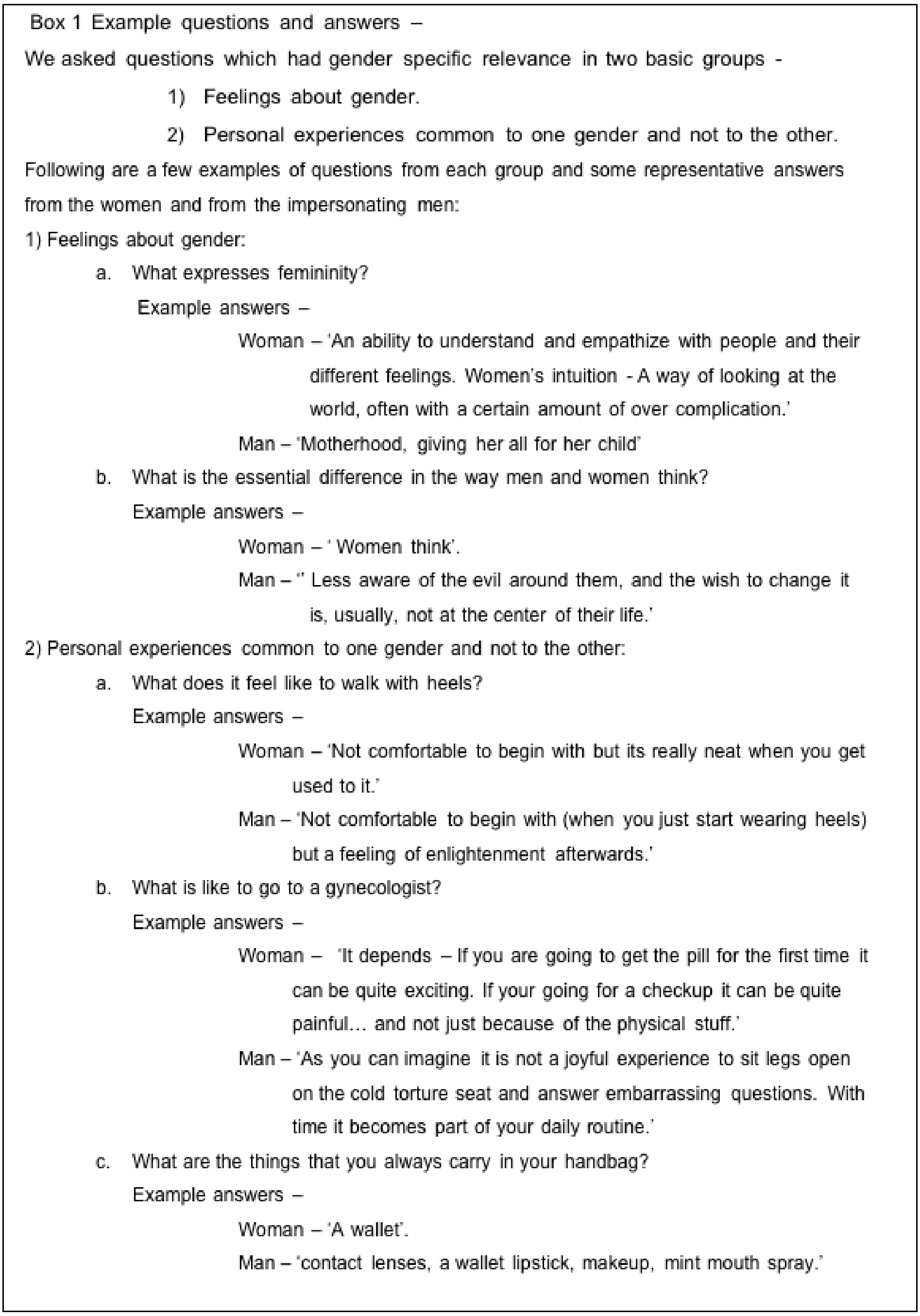
We have finally performed (admittedly 53 year late) the first part of the Turing test. Let us now describe our experiment which is a structured implementation of the game suggested by Turing.

2. Testing the First Part of the Turing Test

Each experimental session included a man (A), a woman (B) and the (female) interrogator. To ensure reproducibility, we used a standard list of 15 questions which were presented in writing simultaneously and independently to the man (A) and to the woman (B). The woman (B) was asked to provide truthful information. The man (A) was asked to give answers that would fool the interrogator into thinking that he was the woman.

The questions were chosen from a larger pool of possible questions by selecting the ones that were found helpful in eliciting answers that identify the man (for examples of the questions and answers, please look in Box 1). In accordance with Turing's requirement that the interrogator can ask as many questions as she wishes, the results of previous experiments were available to the interrogator.[a]

[a] There is a trap of infinite self-reference loop: in an infinitely iterated series of encounters, the strategy of the interrogator may adapt to the interrogated and vice versa. This leads to an infinite loop that is practically untreatable by finite experiments. A reasonable procedure is to have the interrogator acts in a potentially infinite pool of potential candidate to be interrogated. Consequently, each person/entity will be typically interrogated only once and the paradoxes connected to the infinite adaptation loop are avoided: there is no way the interrogator can adapt to a particular individual: he will adapt to the population of women and men as composing its environment. Correspondingly, the task of A (whether man or computer) is to win its single shot with the interrogator, not a recurring series of return matches building one upon the other.



Box 1 Example questions and answers –

We asked questions which had gender specific relevance in two basic groups -

1) Feelings about gender.

2) Personal experiences common to one gender and not to the other.

Following are a few examples of questions from each group and some representative answers from the women and from the impersonating men:

1) Feelings about gender:

a. What expresses femininity?

Example answers –

Woman – 'An ability to understand and empathize with people and their different feelings. Women's intuition - A way of looking at the world, often with a certain amount of over complication.'

Man – 'Motherhood, giving her all for her child'

b. What is the essential difference in the way men and women think?

Example answers –

Woman – ' Women think'.

Man – '' Less aware of the evil around them, and the wish to change it is, usually, not at the center of their life.'

2) Personal experiences common to one gender and not to the other:

a. What does it feel like to walk with heels?

Example answers –

Woman – 'Not comfortable to begin with but its really neat when you get used to it.'

Man – 'Not comfortable to begin with (when you just start wearing heels) but a feeling of enlightenment afterwards.'

b. What is like to go to a gynecologist?

Example answers –

Woman – 'It depends – If you are going to get the pill for the first time it can be quite exciting. If your going for a checkup it can be quite painful… and not just because of the physical stuff.'

Man – 'As you can imagine it is not a joyful experience to sit legs open on the cold torture seat and answer embarrassing questions. With time it becomes part of your daily routine.'

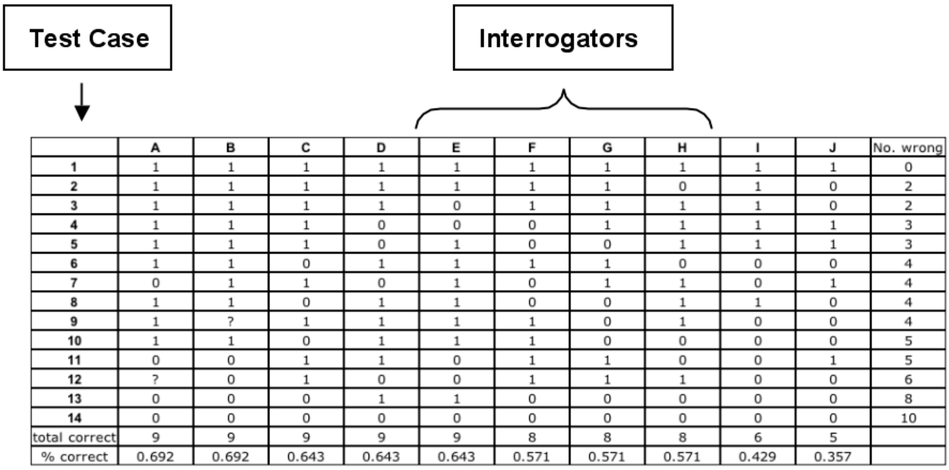
c. What are the things that you always carry in your handbag?

Example answers –

Woman – 'A wallet'.

Man – 'contact lenses, a wallet lipstick, makeup, mint mouth spray.'

Table 1. Levels of success by interrogators in identifying the males in the first stage of a Turing test (i.e., when A is a man). Each column shows the results for a different interrogator (1 — correct identification of (A), 0 — incorrect identification). Each row is a different test case of a pair of subjects. The rows and columns are aligned by how successful the interrogators were. In test case one, the male was always identified, in test case 14, never. Interrogator A was the most successful, interrogator J, the least successful (note — due to technical difficulties interrogators A and B tested only 13 pairs of subjects).


Test Case
Interrogators

	A	B	C	D	E	F	G	H	I	J	No. wrong
1	1	1	1	1	1	1	1	1	1	1	0
2	1	1	1	1	1	1	1	0	1	0	2
3	1	1	1	1	0	1	1	1	1	0	2
4	1	1	1	0	0	0	1	1	1	1	3
5	1	1	1	0	1	0	0	1	1	1	3
6	1	1	0	1	1	1	1	0	0	0	4
7	0	1	1	0	1	0	1	1	0	1	4
8	1	1	0	1	1	0	0	1	1	0	4
9	1	?	1	1	1	1	0	1	0	0	4
10	1	1	0	1	1	1	0	0	0	0	5
11	0	0	1	1	0	1	1	0	0	1	5
12	?	0	1	0	0	1	1	1	0	0	6
13	0	0	0	1	1	0	0	0	0	0	8
14	0	0	0	0	0	0	0	0	0	0	10
total correct	9	9	9	9	9	8	8	8	6	5	
% correct	0.692	0.692	0.643	0.643	0.643	0.571	0.571	0.571	0.429	0.357	

The questionnaires were answered in writing by 14 men and 14 women. An impartial aide coupled the 28 answered questionnaires into 14 test cases, which had one man (A) and one woman (B) each. The test cases were then submitted to ten independent female interrogators. The interrogators were asked to identify for each test case the man (A) versus the woman (B). Our results show that men (A) failed to pass the Turing test (were detected by the interrogators as men) in 58% of the cases on average, i.e., with only 15 questions (rather than the unlimited number allowed by Turing) the interrogators identify the women above chance ($t(9) = 2.02$, $p < 0.05$ — see Table 1). This effect is small but consistent and fits in well with known results from the study of how humans identify lies.[15] This is a striking result: fifty years after Turing, not only is there no computer that passes his test but we find that humans fail it.

What does this mean? Are men not sentient? Can't they think? Or was Turing mistaken in connecting sentience to imitation? The answer to all three of the above questions is probably "No". Turing's challenge to find a machine that can imitate man is as justified as ever. The problem is that humans are not able to perform successfully the task he required from the computers.

Human gender impersonators are identified by the human interrogators at a level above chance. In fact our results are similar to the ones found in the research on identification of cheats and liars.[15] In its present form the Turing test inadvertently tests the interrogator's capabilities to detect lies, rather then the men's

(A) capability to imitate. It is plausible that even a man who knows exactly what a woman experiences, would give himself away by the mere fact that he is lying (about being a woman (B)) and that the interrogator is capable of detecting liars.

To avoid this effect, and fulfill without bias Turing's intention one should ask the machine to make an imitation that does not require it to cheat. The option suggested by the Turing Tournament @ CALTECH[4,5] to have the machines compete for the role of interrogator as well as the interrogated (A) may lead soon to very illuminating surprises. After all, we — humans — have a saying: it takes one to know one.

Acknowledgments

R. Adam would like to thank all her friends who helped in the process of designing the questionnaire. U. Hershberg is a recipient of the Levi Eshkol Fellowship given by the Israeli Ministry of Science. Y. Schul's research is funded in part by a grant from the US–Israeli Binational Science Foundation (BSF). S. Solomon's research is funded in part by a grant from the Israeli Science Foundation.

References

1. A. M. Turing, *Mind* **49**, 433 (1950).
2. S. Sheiber (ed.), *The Turing Test: Verbal Behavior as the Hallmark of Intelligence* (MIT Press, 2003), 336pp.
3. J. Moor, *The Turing Test Elusive Standard of Artificial Intelligence* (Kluwer Academic Publishers, Dordrecht, 2003), 288pp.
4. J. Arifovic, Turing tournament: A method for evaluating models of agent behavior, talk at the *8th Annual Workshop on Economics with Heterogeneous Interacting Agents*, Kiel, May 2003.
5. J. Arifovic, R. D. McKelvey and S. Pevnitskaya, An initial implementation of the Turing tournament to learning in two person games, working paper (March 2003). The Turing Tournament @ Caltech official home page is http://turing.ssel.caltech.edu/.
6. R. M. French, *Cognitive Sciences* **4**, 115 (2000).
7. J. Sundman, *Technology and Business* (2003).
8. S. Harnad, *SIGART Bulletin* **3**, 9 (1992).
9. J. Goldenberg, D. Mazursky and S. Solomon, *Science* **285**, 1495 (1999).
10. Y. Stolov, M. Idel and S. Solomon, *Int. J. Mod. Phys. C* **11**, 827 (2000).
11. E. Adi-Japha, I. Levin and S. Solomon, *Cognitive Development* **13**, 25 (1998).
12. J. Weizenbaum, *Computer Power and Human Reason* (Freeman Press, 1976).
13. K. A. Dewdney, *Scientific American* **18** (1985).
14. A. P. Saygin, I. Cicekli and V. Akman, *Minds and Machines* **10**, 463 (2000).
15. P. Ekman and M. O'Sullivan, *American Psychologist* **46**, 913 (1991).