THE INVERTED TURING TEST: HOW A MINDLESS PROGRAM COULD PASS IT

(In *Psycoloquy* 7(39) turing-test.6.french.)

Robert M. French
Department of Psychology
University of Liege
Liege, Belgium
french@segi.ulg.ac.be

ABSTRACT

This commentary attempts to show that the inverted Turing Test (Watt 1996) could be simulated by a standard Turing test and, most importantly, claims that a very simple program with no intelligence whatsoever could be written that would pass the inverted Turing test. For this reason, the inverted Turing test in its present form must be rejected.

Alan Turing is generally remembered for two seminal ideas, one computational, one philosophical. The first was his invention of a simple, but extraordinarily powerful computing device, now called a Turing Machine (Turing, 1936). The second was his invention of a simple, but extraordinarily powerful test of intelligence, now called the Turing test (Turing, 1950). One of the wonderful things about a Turing Machine is that you can tinker with it all you want and you won't be able to improve its power. Whatever "improvement" you add (e.g., a second tape, non-deterministic control, a second read/write head, etc.), you can always show that the new machine is, in fact, no more powerful — faster, perhaps, but no more powerful — than the simple machine Turing first described in the 1930's.

Remarkably, the same would seem to be true for the Turing test. All attempts to increase the power or sensitivity of the original Turing test, at least all of those of which I am currently aware, can be simulated within the framework of the original Turing test. Watt (1996) touches briefly on this fact in recognizing the criticism that "the inverted Turing test is redundant because all of its power of discrimination is available in the standard Turing test". But rather than focusing on this crucial criticism, he comments only that "... a critically evaluated standard Turing test without a time limit would be sufficient to detect the presence of naive psychology. However, given that humans have all these psychological biases in their ascription of mental states, I doubt whether a truly critical version of the Turing test is psychologically possible without some variation in the test." But what does he mean by "some variation in the test"? He seems to be arguing for a more reliable version of the test — namely, his inverted Turing test. In what follows I will attempt to show that this "inverted" Turing test could not only be simulated by the standard Turing test but, most importantly, would ascribe intelligence to programs that are certainly not intelligent.

In thinking about the Turing test, people often tend to overlook the completely unfettered nature of the questions that may be asked by the interrogator. It cannot be overemphasized that *any* question is fair game. Questions are allowed, for example, whose answers rely on knowledge that is declarative ("What is the capital of Senegal?"), procedural ("Please describe how you would tie your shoes"), or — and this is crucial — subcognitive ("Is *Flugblogs* a good name for a start-up computer company?") (French, 1990). Watt, however, says, "It might be possible, with the current state of the art, to use a simple set of linguistic metrics that would unambiguously distinguish between people and computer systems. *I would regard this as cheatin g*[my italics]." But what if this "simple set of linguistic metrics" could be elucidated by the answers to a number of perfectly reasonable questions, as described, for example, in French (1990)? Is this still "cheating"? And how could one ever determine

the point at which a technique that took advantage of some underlying "simple set of linguistic metrics" was cheating? Stipulating unambiguously what would or would not constitute cheating would prove to be as impossible as stipulating that questions must be limited to particular domains. The boundaries between real world domains are so overlapping and so fuzzy as to make impossible adherence to any such stipulation.

The underlying idea of the Watt's inverted Turing test is as follows.

> Instead of evaluating a system's ability to deceive people, we should test to see if a system ascribes intelligence to others in the same way that people do. . . .by building a test that puts the system in the role of the observer... [A] system passes [this inverted Turing test] if it is itself unable to distinguish between two humans, or between a human and a machine that can pass the normal Turing test, but which can discriminate between a human and a machine that can be told apart by a normal Turing test with a human observer.

Now, if we accept the conclusions of French (1990, 1995) — as Watt apparently does — the issue of a machine passing the standard Turing test will never come up, so we can reduce his description of the inverted Turing test to the following: Two separate rooms can contain either two human beings or, alternately, a machine and a human. There are two types of interrogators: I_H, a human, and I_M, a machine. If I_M is just as good as I_H at determining when the rooms contain two humans or when they contain a human and a machine, we will say that I_M (the machine) has passed the inverted Turing test.

This is indeed an elegant inversion of the standard Turing test. But to see how the standard version of the test can accommodate this inversion, assume that the human-interrogator I_H and the machine-interrogator I_M are put in separate rooms. There is a Super-Interrogator, who does not know which room contains I_M and which contains the I_H. The Super-Interrogator is, of course, permitted to ask any questions whatsoever of I_M and I_H, but to simulate Watt's inverted Turing Test, he must restrict his questioning to the following type of interrogation. He says to I_H and to I_M: "From now on, you may assume that you are the interrogator of a Turing test. Your terminal will be connected to two rooms, which may contain either two humans, or a machine and a human. Your job is to tell me if the pair of rooms contains two humans or a human and a machine." Both interrogators judge the entities (human or machine) in the first pair of rooms to which they are connected. Then they are given a new pair of rooms and make a new judgment, and so on. There is some number of pairs of rooms and both interrogators judge all of them and report their conclusions to the Super-Interrogator. Since the Super-Interrogator knows the real contents (human-human or machine-human) of each pair of rooms, he can compare the respective performances of two interrogators. If their answer profiles are indistinguishable to him, the machine-interrogator will be judged to have passed the inverted Turing test. Our first conclusion is therefore that the inverted Turing test can be simulated with little difficulty by a standard Turing test. (Again, this is analogous to the Turing Machine. The way that variants of the standard TM are shown to be no more powerful than the standard TM is by demonstrating that the new machines can be simulated by the standard TM.)

The next question is whether the inverted Turing test is sufficiently powerful to prevent obviously unintelligent programs from passing it. The answer is that it is not. But before this argument can be made, it is necessary to become familiar with a special type of question — subcognitive questions — that can be used by interrogators in Turing tests and that will allow foolproof unmasking of computers as non-humans, unless the computers had lived life as we humans had. A complete discussion of this technique can be found in French (1990) or, in succinct form, in French (1995). Furthermore, Watt in his present article seems to accept the arguments presented in these papers.

The most important point of French (1990) is the immense (and generally unappreciated) difficulty that anything not having lived life as a human being would have in actually passing the Turing Test. We humans respond very consistently to "subcognitive" questions (i.e., questions that draw on the subconscious structure of our minds), such as, "Would Flugblogs be a good name for a start-up computer company?" — Of course not! — or "Would Flugblogs be a good name for air-filled bags that you could tie on your feet to walk across swamps with?" — Sure, not bad! Humans' answers emerge from a vast set of learned, associative, and largely unconscious influences involving sounds (Which word is prettier, "farfalletta" or "blutch"? Why, exactly?), connotations (Would you like it if someone called you a "trubhead"? Why, exactly? How could this be explicitly programmed into a machine?), pictures, smells, past events, and so on ad infinitum. These influences are produced by our continual interaction with our environment. And subcognitive questions tap into the results of this lifetime of human-environment interaction. In other words, these questions subtly probe our vast, complex and intricately interconnected associative concept and sub-concept networks that we have acquired through experiencing the world. They are precisely the kinds of questions that would unfailingly unmask any computer that had not lived life as we had.

So, for the inverted Turing test, here is what we do. We independently prepare a long list of subcognitive questions. Then we venture out into the same population from which the humans participating in the inverted Turing test will be chosen. We interview a representative sample of the population, asking them a relatively large number of questions (the Subcognitive Question List) like:

"On a scale of 0 (completely implausible) to 10 (completely plausible):

- Rate *Flugblogs* as a name Kellogg's would give to a new breakfast cereal.
- Rate *Flugblogs* as the name of start-up computer company
- Rate *Flugblogs* as the name of big, air-filled bags worn on the feet and used to walk across swamps.
- Rate *Flugly* as the name a child might give to a favorite teddy bear.
- Rate *Flugly* as the surname of a bank accountant in a W. C. Fields movie.
- Rate *Flugly* as the surname of a glamorous female movie star.
- Rate *banana splits* as *medicine*.
- Rate *purses* as *weapons*.
- Rate *pens* as *weapons*.
- Rate *dry leaves* as *hiding places*.
- etc."

The distribution over the population of the answers to each of these questions will constitute the Human Subcognitive Profile.

So, for example, no dictionary definition of *dry leaves* will ever include the fact that piles of dry autumn leaves are wonderful places for children to hide in and, yet, few among us would not make that association upon seeing the juxtaposition of those two concepts. (Is this tapping into some simple linguistic metric? Certainly, the metric born of human experience with the world. Is this cheating? Surely not.) In any event, by surveying the population at large with an extensive set of these questions, we draw up a Human Subcognitive Profile for the population. It is precisely this subcognitive profile that could not be reproduced by a machine that had not experienced the world as the members of the sampled human population had. The Subcognitive Question List that was used to produce the Human Subcognitive Profile gives an interrogator — any interrogator — a sure-fire tool to eliminate machines from a Turing test in which humans are also participating.

Now, let us return to our two interrogators, I_M, the machine-interrogator, and I_H, the human-interrogator. First, consider I_H. She will be able to eliminate all machines from the running using some form or another of her own subcognitive question list. In other words, she will always be able to determine those pairs of rooms in which there is a machine and a person. Can a completely unintelligent machine-interrogator, I_M, do just as well? Yes. All that is required is for a human programmer to equip I_M, the dumbest of programs, with the previously established Subcognitive

Question List, the corresponding Human Subcognitive Profile and a small statistics routine. The program will then test each candidate (i.e., one machine and one person) with its Human Subcognitive List and use its statistics routine to compare the results with its Human Subcognitive Profile. The inevitable divergence that any machine being questioned will have with this Profile will inevitably unmask it. But in this case the unmasking has been done not by an intelligent human interrogator but a computer program containing little more than the canned Subcognitive Question List, the corresponding Human Subcognitive Profile and a statistical analyzer.

Similarly, for the pairs of rooms each containing a human. Presumably, I_H will not be able to tell them apart. And, as before, the machine-interrogator, will present both humans with its canned Subcognitive Question List, will compare the results with its Human Subcognitive Profile, will find that neither individual differs significantly from its Profile, and will, like the human-interrogator, conclude, rightly, that two humans are in the rooms. The performance of the human-interrogator and the (clearly unintelligent) machine-interrogator will be, at least with respect to Watt's inverted Turing test, identical.

Of course, subcognitive question lists shows the Turing Test at its very hardest. But we must assume that Turing test interrogators are always trying to pose the hardest, most penetrating questions possible in their quest to unmask the computer. It is important to note that there is not a unique Subcognitive Question List (with its corresponding Human Subcognitive Profile); rather there are infinitely many such lists and profiles. But the key assumption is that these subcognitive question lists will sample subcognitive space in a random but representative manner. Thus, even though I_H and I_M will certainly not be using identical subcognitive question lists, discrminations based on their respective question lists will very likely be the same, insofar as each list samples subcognitive space representatively. As a result, any entity (a machine or a human from a radically different culture than the originally sampled population) that shows a significant statistical departure from one profile will show a similar departure from the other. In short, whenever I_H detects a significant difference from her subcognitve profile (thereby presumably judging the deviant entity to be a machine), so will I_M.

In conclusion, the inverted Turing Test could be passed by an obviously unintelligent program armed only with a completely canned list of subcognitive questions, the human response profile corresponding to those questions and a means of doing statistical comparisons. Consequently, the idea of an inverted Turing test must, at least in its present form, be rejected.

REFERENCES

French, R. M., (1995) Refocusing the Debate on the Turing Test. *Behavior and Philosophy*. 23(1), 61-62.

French, R. M. (1990) Subcognition and the Limits of the Turing Test. *Mind*, 99(393), 53-65.

Turing, A. (1950) Computing machinery and intelligence. *Mind*. 59(236), 433-60.

Turing, A. (1936) On computable numbers with an application to the Entscheidungs Problem. *Proceedings of the London Mathematical Society*, 42, 230-265.

Watt, S. (1996) Naive-Psychology and the Inverted Turing Test. PSYCOLOQUY 7(14), turing-test.1.watt.