Wichita State University Libraries
SOAR: Shocker Open Access Repository

Susan G. Sterrett

Philosophy

Nested Algorithms and “The Original Imitation Game Test”: A Reply to James Moor

Susan G. Sterrett
Wichita State University

Citation

Sterrett, Susan G. “Nested Algorithms and ‘The Original Imitation Game Test’: A Reply to James Moor.” *Minds and Machines* vol.12, no. 1 (2002): 131-136.

A post-print of this article is posted in the Shocker Open Access Repository:

<http://soar.wichita.edu/handle/10057/10702>

Nested Algorithms and “The Original Imitation Game Test”: A Reply to James Moor

SUSAN G. STERRETT

Department of Philosophy, 201C West Duke Building, Duke University, Durham, NC 27708, USA; E-mail: sterrett@duke.edu

In “The Status and Future of the Turing Test” (Moor, 2001), which appeared in an earlier issue of this journal, James Moor remarks on my paper “Turing’s Two Tests for Intelligence.” In my paper I had claimed that, whatever Turing may or may not have thought, the test described in the opening section of Turing’s now legendary 1950 paper “Computing Machinery and Intelligence” is not equivalent to, and in fact is superior to, the test described in a passage that occurs much later in Turing’s paper (i.e., in Section 5 of Turing, 1950). I’m pleased Moor chose to give such prominence to my point, and very happy to see that he recognized that my claim was a normative one about the superiority of one test over another, rather than a claim about Turing’s intentions. However, as I think the way he describes my point could lead to misunderstandings, I’d like to clarify the points I made. One major clarification is which two tests I am contrasting; another is that the difference in overall structure of the two tests is of philosophical significance.

In my paper, I contrasted the test described in Section 1 of Turing’s “Computing Machinery and Intelligence” with the test Turing described in Section 5 of his paper. I dubbed the test in Section 1 “The Original Imitation Game Test”, and referred to the one in Section 5 as “The Standard Turing Test”. These appear as the two headings in Figure 1 (“The Two Tests and How They Differ”) of my paper (Sterrett, 2000, p. 544). I argued for the superiority of the neglected “Original Imitation Game Test” over the much-discussed “Standard Turing Test”, summarizing three main points in that figure: In the Original Imitation Game Test: “(1) test structure permits the result that the machine does better than the man, (2) test tends to screen off lack of interrogator skill, and (3) both man and machine are required to impersonate. The machine’s performance is not directly compared to the man’s, but their rates of successfully impersonating against a real woman candidate are compared.”

I was careful not to refer to the test described in Section 1 of Turing’s paper as “the gender game”, “the gender test” or “a gender imitation test”, though these are terms other commentators on Turing’s paper have used to make points of their own. In contrast, I explained that although Turing did describe the test using a task involving gender impersonation, the valuable insight reflected in the design of “The Original Imitation Game Test” had to do with the structure of the test. In fact,

I stated that gender impersonation and discrimination were inessential to the test, while explaining why it was a good task to employ in a test of intelligence. I wrote: “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.” I went on to explain:

The significance of the cross-gendering in Turing’s Original Imitation Game Test lies in the self-conscious critique of one’s ingrained responses it requires. And, that the critique has two aspects: recognizing and suppressing an inappropriate response, and fabricating an appropriate one. (Sterrett, 2000, p. 551)

The only reference I made to gender games or tests was in explaining that other commentators who used the term were describing something different from what I called “The Original Imitation Game Test”. I wrote: “Some [commentators] 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.” The whole point of this very brief mention of a so-called gender game or test in my paper was to clarify that I was not drawing a contrast between a test that employed gender determination and one that employed species determination, as these other writers do. There may be some interesting points to be made by examining such a contrast, but the contrast I urged as having philosophical import is the contrast between a test having the kind of structure Turing describes in Section 1 of the paper, and the test he describes in Section 5 of the paper.

Thus I worry about misunderstandings that could arise from Moor’s well-meant but not quite accurate statement in which he lumped me in with those writers whom I stated had extracted a different test than I had. Moor wrote: “a number of writers suggest that Turing intended or should have intended the gender interpretation (Genova, 1994; Hayes and Ford, 1995; Sterrett, 2000; Traiger, 2000). “It’s not really accurate to describe my point as arguing for “the gender interpretation”. However, I think I can see why Moor did so: this statement of Moor’s occurs in his discussion of two different interpretations of Section 1 of Turing’s paper, which Moor calls the gender interpretation and the human interpretation.

But, on my view, the issue is not a matter of using a “gender interpretation” in contrast to a “human interpretation”. I chose the name “The Original Imitation Game Test” — a name I do not believe was used prior to my doing so – to emphasize the two most distinctive features of the test described in Section 1 of Turing’s paper.

- One distinctive feature is the “nested algorithm” structure of the test. The Original Imitation Game is played repeatedly in the course of a single test, and is nested inside the test. The test is a practical test, which, if passed, would be evidence for regarding a machine’s behavior as intelligent. The game is not the test; the game is nested inside the test, as a DO or DO WHILE loop is nested in an algorithm.
- The second distinctive feature I meant to emphasize in choosing the name “The Original Imitation Game Test” is the kind of game that is nested in the test: the Original Imitation Game sets a candidate the particular task of impersonation and involves a non-candidate who is of the kind the candidate is to impersonate. Winning a round of the game means pulling off an impersonation successfully when pit against a particular non-candidate who is the genuine article, as required by the structure of the Original Imitation Game. Passing the test means being “as good as a man” at winning such rounds. In The Original Imitation Game Test, the candidate machine is never in the same round of an imitation game as a man; each candidate man or machine has the task, to be achieved in his or its own way, of having the interrogator (wrongly) conclude, on the basis of the conversational exchanges carried out in one specific round of the game involving one other particular individual of the sort to be impersonated in addition to the candidate, the following: the candidate, rather than the genuine article, is the thing impersonated, and the genuine article is merely an impersonator. The rules of the game require that the interrogator choose one (and only one) of the two as the impersonator and one (and only one) as the genuine article.

The nested algorithm structure of the “Original Imitation Game Test” gives the first test a significantly different character than the “The Standard Turing Test”. In my paper, I pointed out that the difference in structure has the generally unrecognized but stunning consequence that “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 succeeds in getting the interrogator to say he is B (the woman) 1% of the time, and the machine gets the interrogator to do so 3% of the time. In contrast, the Standard Turing Test does not admit of such a result.” (Sterrett, 2000, p. 545) Moor redescribes my point in terms I did not use: “Sterrett is correct that the existence of a control group in the gender imitation test, compared to the absence of such in a human imitation test, offers a specific standard for comparison.” (Moor, 2001, p. 79) and he proceeds to dismiss the difference between the two tests as of no philosophical consequence on the basis of this misdescription. Actually, I did not talk about “the existence of a control group” or use the terminology of controlled experiments. In contrast, when I spoke of a “suitably chosen sample”, the metaphor I used was taken from baseball; I wrote: “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.” (Sterrett, p. 554)

Although Moor meant to be helpful in redescribing this difference in structure in the contemporary terminology of the methodology of controlled experiments, his attempt to capture the difference in structure between “The Original Imitation Game Test” and “The Standard Turing Test” as a matter of the existence or non- existence of a control group is not an accurate description of the difference to which I meant to be drawing attention. By putting the point this way, one may miss the significance of the nested algorithm structure of “The Original Imitation Game Test.” It seems to me that Moor at least does so, in then going on to argue that there is no advantage to the Original Imitation Game Test, and, especially, in claiming that “the aspects of intelligence that Sterrett identifies as important to test can be tested in the standard game.” (Moor, p. 80)

The first basis on which Moor dismisses the difference is the reiteration of a common complaint often brought against the so-called Turing Test, that “Machines might do as well as ... men but might not demonstrate much intelligence.” (Moor, 2001, p. 79) Since he is here stating a complaint I addressed in my original paper, I refer the reader to my earlier paper for arguments as to why this common complaint, though justified against the “Standard Turing Test”, does not apply to “The Original Imitation Game Test”. (Sterrett, 2001, especially pp. 545–552)

The second complaint Moor makes is “Of course, it might be replied that those machines that did imitate women well did show intelligence. But, it is exactly those machines that would be expected to do well in the standard Turing test and this would not show a normative advantage to using the gender imitation test over the standard test.” (Moor, 2001, p. 80) I’ve already pointed out that the contrast I was drawing was not a matter of using gender imitation as opposed to some other kind of imitation. But that’s not the only reason I don’t find the complaint legitimate. Why should we think that it is “exactly those machines” that do show intelligence that will do well in the Standard Turing Test? Moor’s reasoning here, if I read him correctly, starts from the assumption that any machine that did well on the test I urge is better would also do well on the Standard Turing Test. But, all that follows from this assumption is that passing the standard test is a necessary (though not a sufficient) condition for passing the test I urge is better. Moor goes beyond this when he concludes that the Original Imitation Game Test, which I urge is better, has no normative advantage over the standard test. For, his conclusion doesn’t follow: just because passing test

A is a necessary condition for passing test B does not mean that test B cannot be better than test A! If test A is not sufficiently discriminating, test B may be better in spite of the fact that anything that passes test B would also pass test A! And, in fact, there are many commentators who complain that the Standard Turing Test is too weak a test, citing cases of people being fooled by clever programming tricks, and pointing out that the standard test is too sensitive to the skill and/or gullibility of the interrogator.

In my paper, I argued that the difference in the structure of the two tests I was contrasting has the effect of screening off the particularities of the interrogator, and thus, that, even on the assumption that it makes sense to talk of “passing” a test that is so sensitive to interrogator particularities, passing the standard test is not a sufficient condition for passing “The Original Imitation Game Test”. Thus Moor’s claim that the difference in the two tests is not significant on the basis that passing the standard test is a necessary condition for passing the Original Imitation Game Test ignores the concern expressed by many that the standard test might not ferret out the merely cleverly programmed machines we would not want to regard as intelligent. Such critics are calling for a test that requires more than the standard test does.

My own view is that the “Standard Turing Test” is just too sensitive to the skill of the interrogator to even be regarded as a test. However, putting that issue aside, the objection that the standard test is all-encompassing because it allows one to ask any question one likes makes about as much sense as saying that a tennis tournament is no better a test of an athlete’s ability to respond flexibly and intelligently than a general test characterized as a setup in which a judge observes an individual (or individuals) who are to respond to requests made by the judge, on the basis that one of the requests the judge could make would be to request that the candidate or candidates pantomime playing tennis. It makes a difference how the non-candidates and the other candidates are involved in the process of comparison. The metaphor of batting averages is useful here, too: the overall structure of the setup within which batting averages are calculated involves each player being at-bat against the same, or similarly skillful, pitchers as the other batters against whom he or she is being compared. The pitcher, however, is free to use different kinds of pitches on different batters. No batter’s style or manner is being directly compared against another batter’s style or manner, nor are their individual attempts to hit the ball judged as to the level of skill exhibited. The batting average arises from the multitude of times that the batter is at-bat. Each time at-bat, the batter is to accomplish, in his or her own way, success, i.e., getting a hit. The structure is not exactly like the Original Imitation Game Test in all its details, but the point is this: the reader, I hope, will agree that the following setup is not as good a measure of batting ability as the current method of computing batting averages: A pitcher pitches to two players (attired in costumes hiding their identities, if need be) and judges which player makes the better attempt to hit the ball.

It’s important to realize that a round of the game which is nested within “The Original Imitation Game Test” involves a specific individual of the sort to be impersonated. It’s a little like the superiority of the method of comparing an artist’s rendition of an object with which one is familiar side-by-side with the object itself, over the method of merely seeing an artist’s rendition of that object and having to say whether it’s a good rendition or not. I believe Moor misses this kind of difference, which I consider very important, in arguing that the standard test is flexible enough to test anything that the Original Imitation Game Test can test for. Moor writes: “For example, an interrogator could ask, after the gender roles of A and B had been established, that A and B assume genders opposite their own and answer questions accordingly ... various skills, from imitating the opposite gender to creating poetry to designing a house, could be evaluated within the framework of a standard Turing test.” (Moor, 2001, p. 80) Since Moor is

arguing for what he calls “the human interpretation”, on which A is the candidate machine and B is a human, I am not sure what Moor intends the task of the machine to be here. However, let us put that issue aside for the moment. Whether Moor means that the machine and the human are to give side-by-side impersonations of something they are not, or whether they are to individually give impersonations of something they are not, I trust the reader can see why I say that the game-nested-inside-a-test structure, in which comparison is forced by the requirement that the interrogator must say that one is the impersonation and the other the genuine article, is superior to such a suggestion. Moor’s suggestion is a setup in which an interrogator asks one or more candidates to do an impersonation, and the interrogator, knowing it is an impersonation beforehand, and in the absence of an example of the genuine article in that context, judges how good an impersonation it is. More discussion of the philosophical significance of the overall structure of the “Original Imitation Game Test” can be found in my paper (Sterrett, 2000), should the reader remain unconvinced.

In this reply to Moor, I have focused on showing that there is a significant difference in the structure of the two tests. In the original paper to which Moor was referring, I suppose I thought that, by showing that the two tests give different quantitative results, the issue of the non-equivalence of the two tests would be settled. It has not turned out that way; a number of people have responded (in personal communication) to my paper as Moor has, by defending the conviction that the two tests are not different in any way that makes a difference philosophically. I am glad for the chance to defend the point that the neglected test deserves more attention, and have here been concerned to defend the point that the two tests are not equivalent in service of that goal.

However, I actually consider the points about culture and cognitive habit more relevant to the topic of the nature of intelligence than the points about overall test structure. My conclusion in the original paper was that the Original Imitation Game Test is superior because “it takes a longer view of intelligence than 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.” (Sterrett, 2001, p. 556) Since the publication of “Turing’s Two Tests for Intelligence”, I have come across a number of writings by philosophers about the relation of habit and intelligence, many of which have reinforced this insight. I expect to discuss that aspect of “The Original Imitation Game Test” in more detail in another paper (Sterrett, 2002, forthcoming).

References

Moor, James H. (2001), ‘The Status and Future of the Turing Test’, *Minds and Machines* 11, pp.77–93.

Sterrett, Susan G. (2000), ‘Turing’s Two Tests for Intelligence’, *Minds and Machines* 10, pp. 541–559.

Sterrett, Susan G. (2002, forthcoming) ‘Too Many Instincts: Contrasting Philosophical Views on Intelligence in Humans and Non-Humans’, JETAI.

Turing, A. M. (1950), ‘Computing Machinery and Intelligence’, *Mind* 59, pp. 443–460.