Minds and Machines (2020) 30:469–486
https://doi.org/10.1007/s11023-020-09543-6

GENERAL ARTICLE

The Genius of the 'Original Imitation Game' Test

S. G. Sterrett[1]

Received: 13 May 2020 / Accepted: 30 September 2020 / Published online: 29 October 2020

Abstract
Twenty years ago in "Turing's Two Tests for Intelligence" I distinguished two distinct tests to be found in Alan Turing's 1950 paper "Computing Machinery and Intelligence": one by then very well-known, the other neglected. I also explained the significance of the neglected test. This paper revisits some of the points in that paper and explains why they are even more relevant today. It also discusses the value of tests for machine intelligence based on games humans play, giving an analysis of some twentieth century TV game shows and how they relate to the tests for machine intelligence in Turing's paper and in some other tests for machine intelligence that have been proposed since. Their value in distinguishing between 'wise' and simply 'clever' AI is discussed.

Keywords Intelligence · Machine intelligence · Turing Test · Artificial intelligence · Gender · Alan Turing · IBM Watson

1 Introduction

"The 'Original Imitation Game' Test" was coined over 20 years ago, in the article "Turing's Two Tests for Intelligence", published in this journal (Sterrett 2000) and in talks prior to that.[1] The article in which it was coined showed that there were actually two distinct and very different tests one could locate in Turing's 1950 "Computing Machinery and Intelligence" (Turing 1950): "The 'Original Imitation Game' Test" and the one more commonly taken from it, which I dubbed "The Standard Turing Test." It related hard-won but easily verifiable observations about the nature

[1] CAP'99 (Computers and Philosophy 1999, August 6th, 1999, Carnegie Mellon University, Pittsburgh, PA. https://www.iacap.org/conferences/past-conferences/cap-1999-at-carnegie-mellon/). Also presented at the Pittsburgh Group on Theoretical Cognition and at Occidental College in early 2000. Later papers on the topic have appeared (Sterrett 2002, 2012, 2017).

✉ S. G. Sterrett
susangsterrett@gmail.com

[1] Department of Philosophy, Wichita State University, 1845 Fairmount Street, Wichita, KS 67260, USA

of each of the tests, i.e., about the kinds of quantitative results each could yield, and laid out the virtues and vulnerabilities of each of these two tests—quite objectively, I thought—and showed that "The 'Original Imitation Game' Test" had many virtues and, in contrast, that "The Standard Turing Test" had many vulnerabilities. It then explained why these differences between them were so significant. The 'Original Imitation Game' Test' delivers on promises that the "Standard Turing Test" does not, and is immune to many of the criticisms that the "Standard Turing Test" is not.

In this paper, I want to focus on what it is about the "Original Imitation Game' Test that makes it an appropriate practical means of addressing the question as to when it might make sense to say a machine appears to be thinking. In the title of my contribution to this special journal issue, I speak of the 'genius' of the 'Original Imitation Game' Test in the sense genius is used when speaking of the genius of an institution or law, where it has the sense of "general intent or meaning; characteristic method or procedure." By the 'Original Imitation Game' I mean the specific 'imitation game' described in the first section of Turing's 1950 paper "Computing Machinery and Intelligence' in the journal *Mind* under the heading 'The Imitation Game." Turing uses that game to construct a test; the title "The Genius of the 'Original Imitation Game' Test" reflects that my topic is the general intent of that test, the 'Original Imitation Game' Test. But a double meaning to 'genius' is at play here, too: there is also something genius-like about the proposal to base a test on the 'Original Imitation Game,' in the more usual sense of genius we use, too: as displaying "instinctive and extraordinary capacity for imaginative creation, original thought, invention, or discovery.

The "Turing's Two Tests..." paper has been cited a fair number of times, and is often mentioned, and occasionally even read, in college courses. However, the authors of many publications on it misstate what is said in the paper even when they cite it, sometimes egregiously so.[2] I explicitly stated in that paper that I was not making any claims about which of the two tests Turing "meant" to be presenting in the paper, nor to profess to know what he had been thinking. I separated off the points about the two distinct tests from such historical questions, in such a way that Turing's failure to appreciate that his paper described at least two distinct tests was not relevant to the points I made there about their differences.

"Turing's Two Tests..." was also explicit that gender was not essential to the structure of a test that had the same virtues as "The 'Original Imitation Game' Test,": "… 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." It then explained what it was about cross-gendering that set the appropriate demands in the game, to further emphasize the point that the value of the game did not lie in the cross-gendering per se, but in "the self-conscious critique of one's ingrained cognitive responses" that cross-gendering typically requires. (Today I would qualify this by speaking of what such a critique

[2] As I write this, the Wikipedia entry on the Turing Test says that I conflate the two tests. Whereas, the whole point of Sterrett (2000) is that there are two distinct tests; it even gives them proper names. Other authors similarly describe Sterrett as saying the exact opposite of what Sterrett (2000) actually said about the role of gender in Turing's article, or describe it as making a historical claim about what Turing meant, which is, again, the exact opposite of what that article said.

"requires of cis-gendered participants.") I further explained that "the critique has two aspects: recognizing and suppressing an inappropriate response, and fabricating an appropriate one." And, that "The 'Original Imitation Game' Test" made the fine distinctions between performances of intellectual skill and cognitive habit, and 'that what I called the "Standard Turing Test"[3] did not.

These and other comparisons and novel observations about "The 'Original Imitation Game' Test" made in that paper that I consider most important do not seem to have made their way into most of the philosophical discourse on artificial intelligence yet. Few discussions in the literatures have paid attention to the detailed analysis provided in that paper proving that the two tests described in Turing's paper are in fact different, and in what ways they differ, ("The two tests and how they differ", Fig. 1 of Sterrett 2000, p. 544) nor to the substantial points in the paper about why these differences are significant and helpful for the future of AI.[4] Hence the topic is not exhausted, and in fact it seems to me that the points in that paper (Sterrett 2000, 2002) are more relevant now than ever. Thus, the need to explain their current significance and make the points clearer provides the motivation for writing an essay about it for this special issue of *Minds and Machines*.

2 Why a Game and Not Just a Conversation?

In one of the papers on the topic of Turing and machine learning and intelligence published in the intervening 22-odd years (Sterrett 2012), I further developed a point in the "Turing's Two Tests..." paper: the importance of appreciating the game context in understanding the structure of the tests. The conversations in both "The 'Original Imitation Game' Test" and "The Standard Turing Test" should not be seen as merely conversations of the sort that might take place in casual conversation, as is so often portrayed in the philosophical literature. Rather, they occur in game set-ups with protocols and time limits, where specific roles and the goals associated with them matter in evaluating language performances. The game context provides means to hone in on the part of language performances that have to do with being reflective and resourceful, i.e., not 'machine-like.'. This is important, since many of our utterances in normal conversation *are* machine-like in that they are made out of habit or convention and so, in much of normal conversation, a response can be appropriate even though it doesn't require much intellectual effort to compose it.

The recognition that the contexts for the proposed evaluation of machine performances are games might seem to trivialize their value in AI research. But, if we are

[3] "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." (Sterrett 2000).

[4] E.g., Shah and Warwick (2016).

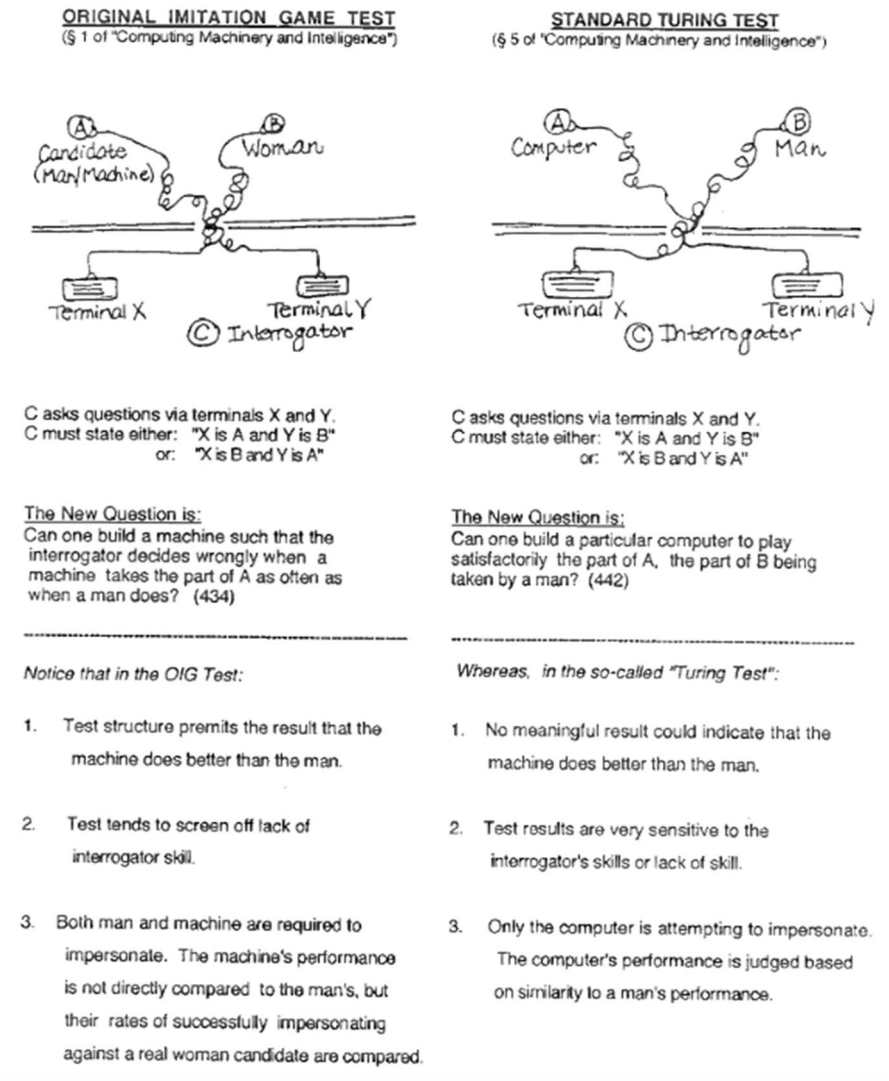

Fig. 1 The two tests and how they differ (from Sterrett 2000)

in fact going to have machines and artificial intelligence programs that interact with humans, we need to be sure we understand what performances in game contexts really show—for the purpose of dismantling hype about AIs winning games against humans, as well as for the purpose of understanding promising avenues for AI to be genuinely helpful, and, above all, not harmful.

The history of AI has been marked by exhibitions of machines competing against humans in games humans already play, since the first electronic digital computer was built: i.e., contests to determine if a computer can beat a human at G, where G is some specific game humans play with each other. As time went on, they progressed

to whether a computer could beat a human player who was a champion at G. One could be skeptical of the purpose and value of these competitions, but not all in the profession feel that way. The IEEE's biography of Arthur Samuel, who was awarded the Computer Pioneer Award in 1987, argued for the value of his work on building programs that could compete with humans in games they already played: "Programs for playing games often fill the role in artificial intelligence research that the fruit fly Drosophila plays in genetics. Drosophilae are convenient for genetics because they breed fast and are cheap to keep, and games are convenient for artificial intelligence because it is easy to compare computer performance with that of people.".[5]

We saw that Turing discussed displays of machine intelligence (in the subjective sense in which he used the terminology) in terms of a computer's ability to play chess. Early on, the challenge set was simply that a machine could play the game of checkers/draughts competently—until Arthur Samuel's own checker-playing program beat him. Then, in 1961, at a publicized event, his program beat the Connecticut State Checker Champion. Samuel's analysis of what it took to improve the program has been recognized as revealing valuable insights into the practice of computer programming and machine learning.[6] Next was Chess. In 1997, IBM's Chess-playing program *Deep Blue* won a tournament with the world chess champion (Garry Kasparov). The kind of programming methods in use were by that time very different from those used in the first checker-playing programs. However, that effort, too, which was massive and marked by failures at first, is now recognized as at least potentially valuable for more general insights about the role of heuristics in artificial intelligence programs. Heuristics that were developed from analyzing human performances were used in developing *Deep Blue*. Many would have liked to see more *Deep Blue* performances before concluding too much from that tournament, but *Deep Blue* was retired shortly afterwards.

Then, in a move that was intended to be a step closer towards a machine that could take on the challenge of a test involving language such as some form of the Turing Test, IBM chose the game of *Jeopardy!* as the next "Grand Challenge." In a televised tournament against the top two human players in the world in 2011, IBM's *Watson* won. The specific content used for the *Jeopardy!* questions and answers, which closely tracked the structure and crowd-sourced content of Wikipedia, had a lot to do with Watson giving such an impressive performance and, even, being able to pull off a win. I discussed what we ought to make of that, in "Turing on the Integration of Human and Machine Intelligence" (Sterrett 2017). What calls into question many of the pronouncements that were made based on Watson's win is that (a) Watson had 'read' all of *Wikipedia* (even though it did not have access to the internet during the tournament), and (b) that Watson's designers had noticed during training Watson to play *Jeopardy!* that over 95% of the correct responses in *Jeopardy!* were titles of *Wikipedia* articles. Of course they used that observation about the game in designing Watson's algorithms (Chu-Carroll 2012). It turned out to be important to understand what Watson's

[5] Biography of Arthur Samuel. *IEEE Computer Society.* https://www.computer.org/profiles/arthur-samuel.

[6] https://www.computer.org/profiles/arthur-samuel

successes were really attributable to, for what happened after that was problematic if not tragic: some at IBM took Watson's capabilities to be more general than they were, not appropriately taking into account the special nature of correct responses in *Jeopardy!* in conjunction with Watson's training on the crowdsourced Wikipedia articles it had 'read' in preparation for the tournament had had in Watson's success. In the wake of Watson's success, the marketers made big plans for Watson: Watson was going to be an MD! Or, as close to being an MD as an artificial intelligence program can be.

When applied to medical diagnosis, Watson's abilities to make appropriate diagnoses based upon what it 'read' in medical journals or was fed by its trainers was far less successful than its glowing performance on the set of *Jeopardy!* had led customers to believe (Ross 2018; Strickland 2019). Understanding how machine performance on one task relates to performance on another is important. Analyzing what is involved in a machine succeeding at some games that humans play should help with that, if done correctly. Once the 'trick' Watson used to win is known, some aspects of its performance in the *Jeopardy!* tournament begin to look more like a demo under controlled conditions than a true competition of question-answering abilities. Yet other aspects of its performance, such as its ability to parse and decode difficult clues, remain impressive. It is worthwhile understanding how to properly analyze the structure, virtues, and vulnerabilities of the games that makers of artificial intelligences will be having their creations play as exhibitions of their capabilities.

Most recently, in 2017, another company took on a 'Grand Challenge" of the game of Go against the world's best player, and its program AlphaGo won. In turn, a successor of AlphaGo, AlphaGo Zero, soon beat AlphaGo (Silver et al. 2017). Then came AlphaZero, which used convolutional neural networks and could play chess and Shogi ("a Japanese version of chess") as well as the game of Go. In fact, it could beat any of the programs specialized to play any of these games. Its creators regarded it as "a notable step toward achieving a general game-playing system." (Silver et al. 2018).

What should the next challenge be? In previous work, I suggested that an appropriate 'Grand Challenge' to take the place of "The Turing Test" (by which is usually meant, "The Standard Turing Test") would be for a computer to be on "To Tell the Truth" playing against humans. "To Tell the Truth" provides a good example of a game that has the same structure and virtues as the 'Original Imitation Game' Test, but does not involve gender impersonation.

3 "What's My Line?", "To Tell the Truth", and the "Original Imitation Game"

As is the case with the "Original Imitation Game," the iconic television show "To Tell The Truth", which was on television in both the US and the UK, had three different roles for players: the role of interrogator/panelist, the role of 'central character' and the role of imposter, who aims to impersonate the 'central character.' In discourse about the "Original Imitation Game" Test, I've observed that people tend to find it hard to keep a concrete grip on how Turing's 3-person 'imitation game'

is structured, what roles the human–machine distinction and the human–woman distinction have, and how the game is played. The details and significance of the 3-person setup seem to slip through a discussant's cognitive grasp once discussion includes other setups for evaluating machine intelligence, or the focus shifts from the role of the interrogator to the role of the impersonator. In this paper, I will try to counter this perennial problem by first presenting the simpler game from which "To Tell the Truth" was developed: a show called "What's My Line?" "What's My Line?" premiered in the US on February 2, 1950, in the same year that Turing's "Computing Machinery and Intelligence" was published (Berman 2020).

"What's My Line?" did not have a role requiring impersonation. Rather, "contestants were asked simple yes-and-no questions by the panel members,... who tried to determine what unusual or interesting occupation the contestant had."[7] The contestant had to answer truthfully, but could only answer "Yes" or "No." The show's host sat beside the contestant, sometimes conferring with them about whether "Yes" or "No" was appropriate, sometimes clarifying the contestant's answer to the panel. When the answer to a question was "No" the contestant was credited an amount, and the panel member lost an amount. Each panel member got to ask questions until a "No" was received, so there was an incentive to formulate an informative question in a way that would elicit a "Yes." The game ends when the panel correctly guesses the contestant's occupation, or when ten "No"s have been given, whichever occurs first. Some examples of contestant occupations are: a woman who was a plasterer, and a man who was an executive at a diaper service.

Since the contestant can only answer "Yes" or "No", the skills required of the contestant in participating in "What's My Line?" are basic language competency and common sense knowledge about the world. In contrast, the role of panelist provides an opportunity to show cleverness and resourcefulness in composing questions, and in guessing the contestant's occupation based on the answers to all four panel members' questions. The panelists and audience were told the contestant's name and where they lived at the beginning of the show. The audience members (but not the panelists, of course) were told the contestant's occupation before the questioning began, so a great deal of the entertainment value of the show was in watching the panel members struggle to hone in on a productive line of questioning that would culminate in making a correct guess. In summary, there were only two roles on "What's My Line?": the role of member of a panel that interrogated the contestant and tried to infer their identity or line of work, and the role of contestant, who answered 'Yes' or 'No' truthfully to questions asked by panelists.

So, in "What's My Line?", though both roles required basic language competency and common sense knowledge of the world, the role of panelist required a great deal of imagination and resourcefulness in addition, in order to win. Occasionally the contestant was a celebrity, and the panel members were blindfolded and had the task of determining the celebrity's identity. Celebrity contestants did employ some additional skills by disguising their voices. However, given a particular contestant's occupation

[7] Ref: "Game Show 'What's My Line?' Turns 70" by Marc Berman, February 2, 2020 in *Forbes* magazine. Downloaded May 4th, 2020 from https://www.forbes.com/sites/marcberman1/2020/02/02/game-show-whats-my-line-turns-70/#138d10536b11.

or identity, the outcome of the game "What's My Line?" was almost totally dependent upon, and thus a reflection of, the skill of the panel of interrogators.

The show "To Tell the Truth" was later constructed from "What's My Line?" by making the role of the contestant an intellectually challenging one, too. Instead of a single contestant whose task was to answer "yes" or "no" truthfully, the game had a role for two more contestants who were to impersonate the featured contestant. In "To Tell the Truth", the occupation or identity of the featured contestant was no longer something the panel members had to guess. Rather, the panelists were told the identity or occupation of the featured contestant, and instead had to guess which of the three contestants presented to them for interrogation really met the description of having that occupation or identity. A statement composed by the genuine featured contestant, for example "Don Hutchison", was read by the host of the show, and each of the three contestants, including both imposters, would in this case say: "*My* name is Don Hutchison." The featured contestant who actually fit the description had to answer truthfully, but each of the other two contestants were supposed to compose answers they thought would convince the interrogators on the panel that they were the contestant who met the description. This changed the balance of skill involved: the outcome of the new game "To Tell the Truth" that evolved from "What's My Line?" was not a matter solely of the skill of those doing the interrogation, as it had been on "What's My Line?", but had as much to do with the ingenuity and resourcefulness *the contestants who had the task of impersonating* the featured contestant ("main character") as it had to do with the skill of the interrogators. And, unlike in "What's My Line?", the imposters were actively working to thwart the panelists from making the correct identification.

Sometimes, of course, the nature of the featured contestant's occupation or identity made the task much harder than others. If the contestants who were to impersonate the featured character knew very little about the occupation, events or lifestyle of the featured character, their ignorance could be hard to make up for, no matter how clever they were. For example, for the "Don Hutchison" episode, the description was "a professional sponge diver." Further details read at the outset of the game were given: "My crew and I remain at sea for as long as 3 weeks at a stretch from sunup to sunset 7 days a week. We dive in 2 h shifts gathering sponges which grow in water from 18 to 150 feet deep. My base at Tarpon Springs, Florida is home port for the only commercial sponge fishing fleet in the Western Hemisphere. Signed, Don Hutchison."[8]

Despite being provided such details, very few imposters are going to have the knowledge base to be able to give a correct answer to every question that might be asked to determine which of the three contestants is Don Hutchison. When an imposter doesn't know the answer to a question they are asked, the imposter will of course not be *imitating* the genuine Don Hutchison (since the genuine Don Hutchison will give the correct answer), though he will be *impersonating* him. The imposter has to come up with something that does not betray his ignorance. Alternatively, he can try other tactics, such as deflecting the question or redirecting the conversation. In the episode with the sponge diver, when one of the panelists asked

[8] January 13th, 1964 episode. You Tube channel "To Tell the Truth (CBS)" https://www.youtube.com/watch?v=4KJm7JKf5XQ.

the question: "What is the most dangerous enemy you meet when you're diving?" an imposter answers: "Actually the most dangerous is our carelessness."

The interrogator (panelist) pushes on, pressing him to say what *living things* are a danger to him. The imposter, unable to evade the question, comes up with: "Nothing actually is really dangerous... the sharks and the barracuda..., but nothing really." Now here all the imposter needs to do to win (winning is getting the interrogator to think he is the real Don Hutchison) is to satisfy the interrogator asking the question. He need not be correct, since the interrogators are not experts on the topic of sponge diving either. In fact, this rather evasive answer that I personally found laughably desperate was in fact effective in the context of the game show, in that the imposter who composed it got 3 of the 4 panelist's votes. Did giving the answer: "sharks and barracudas... but nothing really" exhibit intelligence? Well, sharks and barracudas are well known stereotypical large fish, so he chose something to weave loosely into his reply, which shows resourcefulness and knowledge of some sort: if not knowledge about the specific subject matter asked about, then at least knowledge about the stereotypes the panelists are likely to hold about sea creatures, and about the kind of answer that would make them drop a line of questioning that could reveal his ignorance.

The other change to the format of "What's My Line?" that led to the "To Tell the Truth" game format was that each panel member voted independently, without consultation with other panel members. This feature of "To Tell the Truth" recognizes the importance of the skill of individual interrogators to the outcome. That interrogator skill can vary highlights the fact that whether a contestant is a successful impersonator, then, depends on the skill of the interrogator. Should that be so? A bit of reflection on what impersonation involves reveals that it should. An impersonation is considered successful if the impersonator passes the muster of whatever examination is being conducted, and that will depend very much on who is interrogating them and passing judgment on their answers. The kind of intelligence called for is not a matter of knowing everything one would need to know to say exactly what the person being impersonated would say in answer to every possible question, but of knowing how the person judging the impersonation will judge what the impersonator says.

In contrast, the skill of an *interrogator* (a panelist) is <u>not</u> a subjective matter: the winnings of the interrogators on the panel are based on something far more objective: whether the panelist managed to figure out which contestant was the genuine 'main character' and which were the imposters.

The (original) 'imitation game' that Turing presents in his paper can be seen as a modified version of "To Tell the Truth": it is modified so that there is only one imposter, and only one person, rather than a panel of four people, are asking questions of the contestants (the imposter and the 'main character.') The panelist/interrogator cannot see or hear the voices of the contestants, but instead asks questions via text and receives the contestants' replies via text. The 'main character' in the "Original 'Imitation Game' Test" is described more generally than a specific person: the 'main character' to be impersonated by the imposter is described simply as "a woman." Instead of contestant number 1, contestant number 2 and contestant

number 3, there are simply "A" and "B", where "A" designates the imposter (man) and "B" designates the woman.

So, in a nutshell: if the panel of interrogators on "To Tell the Truth" were shrunk down from four people to one person, and the group of contestants answering questions was shrunk from one main character and two imposters to one main character and one imposter, and we were to impose the restriction that the interrogator cannot see the contestants, and that the questions and answers be communicated only via text, we'd get the 'imitation game' introduced on the first pages of Turing's 1950 paper. Turing described it as follows:

> "The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either 'X is A and Y is B' or 'X is B and Y is A'. The interrogator is allowed to put questions to A and B thus:
>
> C: Will X please tell me the length of his or her hair ?
>
> Now suppose X is actually A, then A must answer. It is A's object in the game to try and cause C to make the wrong identification. His answer might therefore be.
>
> 'My hair is shingled, and the longest strands are about nine inches long.'
>
> In order that tones of voice may not help the interrogator the answers should be written, or better still, typewritten. The ideal arrangement is to have a teleprinter communicating between the two rooms. Alternatively, the question and answers can be repeated by an intermediary. The object of the game for the third player (B) is to help the interrogator. The best strategy for her is probably to give truthful answers. She can add such things as 'I am the woman, don't listen to him' to her answers, but it will avail nothing as the man can make similar remarks." (Turing 1950 p. 433–434).

Turing called this the 'imitation game' but used the term more generally later. To refer to the particular game described above, I'll use "Original 'Imitation Game'" per the usage in "Turing's Two Tests..." (Sterrett 2000).

4 The Tests Based on the Games: OIG, QTT, and STT

4.1 OIG Test ("Original 'Imitation Game' Test")

The 'imitation game' described a game played amongst humans, and Turing used it to construct a practical test that indicates when it might make sense to say that a

machine is 'thinking." First, Turing took the (Original) Imitation Game, which is played by three humans, and very slightly modified its description so as to allow for one of the roles, the role of 'A', to be taken on by a machine rather than a man. Thus, this game is played with one machine and two humans, rather than with three humans. In each round of the game that the machine plays in the role of 'A', the machine is thus attempting to make the interrogator believe it is the woman of the pair, i.e., to impersonate a woman. It is thus competing with whatever woman is playing the role of 'B' to be picked as the contestant the interrogator C thinks is the woman. Likewise, when a man plays the role of 'A', the man is competing with whatever woman is playing the role of 'B' to make C think he is the woman. Both the machine and the man have the task of impersonating something they are not. They may go about strategizing and performing the task differently, but the task is the same: impersonation. This allows Turing to compare a machine with a man, by setting up some rounds of the game in which the role of A is played by a machine, and some rounds of the game in which A is played by a man. By comparing the percent of times the machine succeeds in the rounds it plays with the percent of times a man succeeds in the rounds of he plays, the machine's performance can be compared to the man's performance, without ever directly comparing their performances. He proposes a test based on such comparisons:

> "We now ask the question, 'What will happen when a machine takes the part of A in this game?' Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman ? These questions replace our original, 'Can machines think?'" (Turing 1950 p. 433–434).

This is what I call Turing's "Original 'Imitation Game' Test."

In "Turing's Two Tests..." (Sterrett 2000), I pointed out that the "'Original Imitation Game' Test" challenges a player in the role of A to recognize bias in their own responses, and to overcome bias in one's responses—for the sake of winning at a game. And looking back now, from the twenty-first century, the skill of recognizing bias in its own learned responses seems an even more fitting test for intelligent vs unintelligent AI than ever. For, bias—especially bias that results from how an AI system is trained—is now recognized as one of the most pervasive and urgent problems in artificial intelligence.

This, I think, is the insight behind using the original 'imitation' game that Turing described at the outset of his 1950 paper. Notice how distinctive the requirement to recognize bias in one's own learned responses is, and then to be resourceful in constructing a response to meet the need at hand from what one has learned. It is not a matter of 'behavioral similarity' to a human, as many seem to assume, and as Paul Churchland once explicitly said that he assumed any test of machine intelligence must be (Churchland 1996). A better name for the 'imitation game' would have been the 'impersonation game.' Thus, *one distinctive feature of the OIG Test is that it requires reflection on one's learned responses, rather than on having learned responses that mimic the responses of something or someone else.*

Another distinctive feature of the Original Imitation Game Test has to do with the kind of outcomes that are possible. It could be that the man doesn't do well at all and gets only 1%. It could be that the computer doesn't do well at all, either, but manages to get more than the man, say, 3%. In that case, then, the computer has done better than the man. It's at least a possibility, on this setup. We can make sense of that, if we note that it makes sense that both the machine and the man had to recognize when their normal responses to a question might give them away and that they needed instead to construct a response that would not give their true identity away. They are doing comparable tasks and there is no logical constraint ruling out that the computer might be able to do better at it than the man. This is not possible on tests based on comparing *how similar* the computer's performance is to a man. *Thus, another distinctive feature of the OIG Test is that it allows for the outcome that the computer outperforms the man.*

4.2 STT (The "Standard Turing Test")

Other distinctive and desirable features of the OIG Test are revealed when it is compared with the other test in Turing's paper, which is the test that many have in mind in speaking of "The Turing Test." Later in his 1950 *Mind* paper, Turing described another test inspired by the 'imitation game', which he said was equivalent to the earlier test I dubbed the OIG Test:

> '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 1950, p. 544).

In "Turing's Two Tests..." I noted that "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.... The test for machine intelligence in this second version is then simply how difficult it is for the 'average' interrogator to correctly identify which is the computer and which is the man. [Turing indicates elsewhere a 5-min time limit for the interrogator to do so.] This is what I shall call the *Standard Turing Test*." (Sterrett 2000, p. 542–543). In light of our discussion above about the game "To Tell The Truth", we can describe the Standard Turing Test as a different variation of it, or as a Masculinized version of the OIG Test. The 'main character' is no longer a woman, but a man.[9] So the role of B is always filled by a man in the STT. The role of A is always filled by a computer in the STT. The interrogator is no longer making a judgment concerning which contestant is which gender, and the man is never required to impersonate. Assuming that the interrogator knows that one contestant is a man and

[9] Colin Allen has pointed out that it is very likely that in using 'man' here, Turing meant no more than to indicate the player was human, of either gender. I find this very plausible.

the other is a machine imposter, the interrogator is determining which is the man. Notice that the interrogator is directly comparing the performance of the man and the machine, to determine which is the man. This is very different from comparing how successful a man's attempts to impersonate a woman are with how successful a machine's attempts to impersonate a woman are.

Notice that the Standard Turing Test, or STT, has neither of the two distinctive features of the OIG Test described above. The reasoning supporting this conclusion is laid out in "Turing's Two Tests..." so I shall not describe those arguments here. But I do want to discuss how the two tests differ on yet another distinctive feature of the OIG Test: the OIG Test tends to screen off sensitivity to the interrogator's skill. This is because, in the OIG Test, the interrogator never directly compares the performances of the man and the machine. Rather, each contestant is separately rated on the ability to make the interrogator think they are the woman. Thus if the interrogator is probing and discerning, neither the man nor the machine would be expected to do well on the OIG Test, on which their task is to impersonate a woman. If the interrogator is easy to fool, then both the man and the machine have a chance to win their round, by successfully being chosen as the woman in the rounds that each (separately) plays.

In the STT, on the contrary, the interrogator is directly comparing the man and the machine. In the STT, the man doesn't have to do anything but respond to questions as he normally would, but of course the machine has a much more demanding task than the man does: the task of impersonating the man. Thus the odds are stacked in favor of the man in the STT. If the interrogator is not very probing or discerning, the computer may have a chance. If not, probably not. Reflecting on the influence of interrogator skill on the outcome of the STT shows the main weakness of the STT: the dependence of the test outcome on interrogator skill. Table 2 shows the relation of the Standard Turing Test to the game from which it is derived: what I call the Masculinized Turing Test. In the middle column of Table 2, the weakness of the STT becomes clear: in the Masculinized Imitation Game, what is rewarded is the Interrogator's ability to distinguish between a man and a machine impersonating a man. So, the game itself is the test; the outcome of the STT just is the outcome of the Masculinized Imitation Game. There's not a contest between man and machine to do a certain task well. Since the judgment the interrogator makes is the very distinction we want the test for, we see the potency of the test reduces to the skill of the interrogator.

Is the OIG Test any different in this regard? Yes, it is. The interrogator's judgment is not itself the test in the OIG Test, and the structure of the test (multiple and separate rounds for the machine and the human) is such that the skill of the interrogator is screened off.

It is desirable that a test for machine intelligence provide a way to compare how well the machine fares in comparison to a human, but we do not want the results to be constrained so that the machine can never do better than a human. Nor do we want the test to be a matter of how similar a machine is to a human. The OIG Test fulfills those criteria; the STT does not. Further, we do not want the test to be a reflection of the skill of the interrogators or judges involved in the test, so we need a test that screens that factor off. Again, the OIG Test fulfills this criterion, whereas

Table 1 Games and tests for machine intelligence based upon them

Game	Roles	Skill rewarded in game	Test based on game	What test indicates
"What's My Line?" (1950)	Two roles: Panelist (4) Contestant (1)	*Panelists*' collective ability to figure out contestant's identity or occupation	The questioning Turing Test (Damasino 2020)	Skill, resourcefulness, and knowledge of *Interrogator/Machine*
"To Tell the Truth" (195?)	Three roles: Panelist (4) 'Main Character"(1) Imposter (2)	*Imposters'* ability to fool panelists (2) *Panelists'* individual ability to distinguish between genuine and imposters matching a description (4)	None so far	
The (Original) "Imitation Game" (per Turing 1950 p. 434)	Three roles: Man = A (1) Woman = B (1) Interrogator = I (1)	*Man's* ability to fool interrogator into thinking that of the two he is more likely the woman (1) *Interrogator's* ability to distinguish between **woman** and **manimpersonating a woman**. (1)	The 'Original' Imitation Game Test" (Turing 1950; Sterrett 2000)	Skill, resourcefulness and 'knowledge' of *Machine* in comparison to *Man*. (Each takes turns playing the role of A.) Test result is comparison of the percent of time Man wins when in the role of A with the percent of time Machine wins when in role of A An individual round of the game reflects a combination of the skill of whoever is in the role of A (*Man/Machine*) and the skill of *Interrogator* in distinguishing between woman and impersonator of a woman In each round either a man or a machine is playing role of A and a woman is playing the role of B. The man and the machine are never being interrogated in the same round of a game

Table 2 The "Standard Turing Test" and the game it is based upon

Game	Roles	Skill rewarded in game	Test based on game	What test indicates
Masculinized "Imitation game" (per Turing 1950, p. 442)	Three roles: Machine C = A (1) Man = B (1) Interrogator = I (1)	*Machine's* ability to fool interrogator into thinking that of the two it is more likely the man than the machine; (1) *Interrogator's* ability to distinguish between **man** and **machine impersonating a man**. (1)	The 'Standard Turing Test" (Turing 1950; Sterrett 2000)	Skill, resourcefulness and knowledge of *Machine* in combination with lack of skill and resourcefulness on part of *Interrogator* Test result is the percent of time Interrogator is fooled by Machine C into thinking it is the Man No skill involved in man playing part of B (Man) In each round Machine C plays role of A and Man plays role of B and Interrogator is directly comparing their performances to each other

the STT does not. As the arguments presented in "Turing's Two Tests..." and summarized in Fig. 1 of that paper (Sterrett 2000, p. 544, reprinted below) establish, the OIG Test and STT are not equivalent (in spite of what Turing assumed), and the OIG Test has the features desired in a test for machine intelligence.

4.3 QTT (The "Questioning Turing Test")

The OIG Test is thus distinct from, and superior to, the STT as a practical test for machine intelligence. But, are there any other tests of linguistic performances we might consider based on our analysis of the games we've looked at, i.e., "What's My Line?", "To Tell the Truth" and the "Imitation Game"? Table 1 summarizes points from our analysis of games and the tests for intelligence based on them, and Table 2 summarizes points from our analysis of the STT and the game it is based upon.

Looking at Tables 1 and 2, there are two more options one might consider tapping into to derive a test for machine intelligence: having a machine play a contestant on the unmodified game "To Tell the Truth", and having a machine play the role of panelist in "What's My Line?" The latter suggestion is very much like a recent proposal presented at the "Rethinking, Reworking and Revolutionising the Turing Test" Conference in 2018, by Damassino, based on his dissertation called "The Questioning Turing Test." (Damasino 2020).

The "Questioning Turing Test" can informally be described as "a twenty-questions game" where the questions are answered with "Yes" or "No", similar to "What's My Line"—except that there is only one panelist (questioner) and the contestant (answerer) is thinking of a famous figure, rather than actually being a famous figure. The machine is to take on the role of panelist/questioner at times, and a human is to take on the role of panelist/questioner at times, with the goal of correctly identifying the famous figure, or celebrity, that the contestant/answerer has in mind. Thus, as on "What's My Line?", given the famous figure, it is the questioner's skill that determines the outcome of the game. One of the dimensions on which the questioner is rated is correctly identifying the celebrity the answerer has in mind, and another is the number of questions it takes to obtain the right answer (Damassino 2020, p. 130). Although Damassino also includes another dimension for how similar the machine's questions are to ones a human would produce ('human-like'), I do not see why that couldn't be dropped to produce a slightly revised version of the QTT. Both the machine and human performances could be scored on just these two dimensions. If they were, then such a slightly revised QTT would, like the OIG Test, not unfairly favor the human, and could provide a practical test of machine intelligence. The results of the test would be a quantitative matter, and the man and machine would never be directly compared; rather, their successes in the game would be compared. Such a test would also permit the result that the machine can beat a human. An additional virtue of the QTT, if the dimension of the 'human-likeness' of the responses is dropped, is that the test result is not sensitive to the skill of a human judge. (This is not to say that there are not some uses for an AI test that includes a dimension for the 'human-likeness' of a response, just that a test for machine intelligence that does not require similarity to a human is more general and has many virtues.)

5 Conclusion

Because so many people refer to the OIG Test as the gender test, or think gender is essential to it having the distinctive features it has, I want to make it clear that the difference between the OIG Test and the STT is not a matter of including gender in the OIG test and leaving it out of the STT. I hope by now the reader will recognize that the structure of the two tests is very different, and that, although reflectiveness on gendering and gender stereotypes is used to good effect in the test, that nevertheless gender is not where the 'genius', or spirit, of the "Original 'Imitation Game' Test" lies.

What, then, is the role of gender in the OIG Test? From the "Turing's Two Tests..." paper? I still endorse what I said in the paper in which I first drew the distinction: "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.... similarity to the man's performance itself is not the standard against which the machine is compared." (Sterrett 2000).

And:

"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." (Sterrett 2000).

Thus, in order to create a task on which machine intelligence in constructing linguistic performances can be fairly evaluated compared to a man's linguistic performances, the cross-gendering task is used. Reflecting on what the man needs to do in taking on the role of A in the OIG Test, we see that the issue is really about learned responses and taking on a role—or putting oneself in a position—where interactions are very different than they were during the learning phases of one's life. By showing how the OIG Test can be seen as a variant of the game show "To Tell the Truth," which in turn was developed from "What's My Line?", we can see the intellectual challenge of cross-gendering one's linguistic responses as a special case of the intellectual challenge of impersonation. Though other aspects of one's identity could be used in constructing a test, gender is an especially appropriate choice, however, due to how ingrained gender is in linguistic responses, as the work of linguist Deborah Tannen (1990, 1994) and others have shown.

In the last decade, there has also been an awareness of the fact that stereotypes about gender, among many other classifications such as class, race, and nationality, are ingrained not only in words, but in images and other cultural products that are used, often without recognizing it, in designing tech products including artificial intelligence programs. The intelligence needed now and in the future includes the kind of self-reflectiveness on our training and our ingrained responses that the OIG Test makes central. It is thus becoming clear that being able to recognize learned responses that are inappropriate and should be overridden is going to be very important to successful AI, or what we would consider 'wise' AI versus simply 'clever' AI. So a notion of intelligence based on being able to impersonate is neither fanciful nor frivolous. It is extremely relevant to building good AI, and there are already applications in dire need of this kind of machine intelligence. Ingrained biases have been shown to occur in AI used by banks and financial institutions, by health care and health insurance providers, and by courts and law enforcement. Concerns about bias that arise from machine learning are now mainstream, especially after the appearance of bestselling books on the subject (O'Neill 2016; Noble 2018; Benjamin 2019, esp. Ch. 2). Many more studies, papers, and books have followed.

I hope my comments here have made the points I was concerned to get across in "Turing's Two Tests..." clearer, and that I have shown why I think they are important in helping us to get a grasp on how to evaluate machines that we build with the intent of being helpful. I offer this small contribution about notions of machine intelligence in the hope of a future for good AI.

References

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. Cambridge: Polity Books.

Berman, M. (2020). Game show 'What's My Line?' turns 70. February 2, 2020 in *Forbes* magazine.

Chu-Carroll, J. et al. (2012). Finding needles in the Haystack: search and candidate generation. *IBM Journal of Researchand Development* 56(¾), Paper 6 May/July 2012.

Churchland, P. A. (1996). Learning and conceptual change: The view from the neurons. In A. Clark & P. J. R. Millican (Eds.), *Connectionism, concepts and folk psychology: The legacy of alan turing* (Vol. 2). Oxford: Clarendon Press.

Damasino, N. (2020). *The questioning turing test.* Ph.D. Dissertation, Philosophy. The University of Edinburgh.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York: NYU Press.

O'Neill, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Crown Publishers.

Ross, C. (2018). IBM's Watson Supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. Retrieved July 25, 2018 from https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/

Shah, H., Warwick, K. (2016). Imitating gender as a measure for artificial intelligence: Is it necessary? In *Proceedings of the 8th International Conference on Agents and Artificial Intelligence (ICAART2016)* (pp. 114–119).

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science, 362*(6419), 1140–1144.

Silver, D., Schrittwieser, J., Simonyan, K., et al. (2017). Mastering the game of Go without human knowledge. *Nature, 550*, 354–359. https://doi.org/10.1038/nature24270.

Sterrett, S. G. (2000). Turing's two tests for intelligence. In *Minds and Machines,* Vol. 10, pp. 541–559. Reprinted in *The Turing Test: The Elusive Standard of Artificial Intelligence*. Edited by J. H. Moor. Kluwer Academic, 2003. (Preprint: https://philsci-archive.pitt.edu/8480/)

Sterrett, S. G. (2002). Too many instincts: Contrasting philosophical views on intelligence in humans and non-humans. *JETAI* (*Journal of Experimental and Theoretical Artificial Intelligence*), *14*(1), 39–60. Reprinted in *Thinking About Android Epistemology*, Edited by K. Ford, C. Glymour & P. Hayes, MIT Press (March 2006).

Sterrett, S. G. (2012). Bringing up turing's child-machine. In *How the World Computes*, *Springer Lecture Notes in Computer Science*, 2012, Vol. 7318/2012, 703–713. (Preprint: https://philsci-archive.pitt.edu/9085/)

Sterrett, S. G. (2017). Turing and the integration of human and machine intelligence. In J. Floyd & A. Bokulich (Eds.) *Philosophical explorations of the legacy of alan turing: Turing 100*, *Boston studies in the philosophy and history of science*. Springer, Switzerland, 2017. (Preprint: https://philsci-archive.pitt.edu/10316/)

Strickland, E. (2019). IBM Watson, heal thyself. *IEEE Spectrum, 56*, 24–31.

Tannen, D. (1990). *You just don't understand: Women and men in conversation*. New York: William Morrow and Company.

Tannen, D. (1994). *Gender and discourse*. Oxford: Oxford University Press.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *LIX*(236), 433–460