# Criticisms of the Turing Test and Why You Should Ignore (Most of) Them

Katrina LaCurts
katrina@csail.mit.edu

MIT CSAIL, 6.893

# 1 Introduction

In this essay, I describe a variety of criticisms against using The Turing Test (from here on out, "TT") as a test for machine intelligence. These criticisms range from easy-to-refute to much more complex, and force us to examine the subtleties of the TT as well as our own intelligence. While many of these criticisms make very deep and interesting points, I argue that most of these criticisms are, at this point in time, irrelevant. Though I do not necessarily believe that the TT will forever and always be the best test for intelligence, my main thesis is as follows: before the TT can be changed in any meaningful way, we need to understand *much* more about not just intelligence in machines, but intelligence in humans, and one of the best ways to do so is to work towards building machines that could pass the TT as it stands today.

# 2 The Turing Test

Before discussing any criticisms of the TT, let me first describe the test itself. In [18], Alan Turing put forth his "Imitation Game" as a means to sidestep the question of "Can machines think?" by giving AI a more precise goal. In this game, a human interrogator converses with two separate entities in locked rooms: a human and a machine. The interrogator's goal is to determine which entity is the machine while the machine and the human both try to convince the interrogator that they are human.[1] Turing believed that the goal of AI should be to create machines that can pass this test, i.e., that are linguistically indistinguishable from humans.

# 3 Subtleties of the Turing Test's Output

The first criticisms that I will address are minor, and deal only with subtleties of the machine and interrogator's output. However, they begin to motivate our thinking about the deeper implications of the TT.

## 3.1 The Turing Test Encourages Mistakes

Since a machine must be virtually indistinguishable from a human in order to pass the TT, it must inevitably make mistakes. Turing himself indicated exactly this in his original paper:

> "The machine (programmed for playing the game) would not attempt to give the right answers to the arithmetic problems. It would deliberately introduce mistakes in a manner calculated to confuse the interrogator."

Here, Turing is speaking specifically of arithmetic problems, but we can extend his logic to other trivial types of mistakes. A second type of "mistake", which Turing probably did not foresee, is the machine giving too articulate an explanation for something that humans do intuitively; this is referred to in [9] as "superarticulacy".

In [9], Michie criticizes the TT for encouraging machines to make both types of mistakes: "surely one should judge a test as blemished if it obliges candidates to demonstrate intelligence by concealing it." However, we can resolve this criticism in one of two ways. First, it is possible that mistakes will come about as a result of the machine's intelligence algorithm. After all, humans

[1] We should assume this interrogator is well-trained. It is actually surprising to see how poor some humans are as interrogators; see transcripts from the Loebner contest as reported in [14].

generally do not make mistakes intentionally; it happens as a result of the internal workings of our brain (we forget certain facts, etc.). In this case, the TT is fine as it stands. Second, if we get to the point of creating machines that could pass the TT but for the fact that they make too few mistakes, the "mistake module" would almost surely be trivial to add in. If it were not, that would be an interesting result in its own right, and certainly those machines should be deemed intelligent and the TT modified. But note that we have to get to the point of almost passing the *current* TT before that is an issue.

#### 3.1.1 Superarticulacy

The superarticulacy mistake brings up an interesting idea that is generally not addressed by criticisms of the TT: using the machines as part of a positive feedback loop with humans. On the road to developing intelligent machines, superarticulacy will probably not appear out of the blue. More likely, we will first develop a machine that is only slightly more articulate than humans regarding one or two skills that humans do intuitively. As a result, this machine will give us insight into our *own* intuitive skills, and we will be able to articulate them as well as the machine did. The point is that as we build intelligent machines, they will help us become more intelligent about ourselves. This is excellent motivation for pursuing the TT as a means to learn more about intelligence.

### 3.2 The Turing Test Doesn't Provide a Gradient of Intelligence

The next criticism of the TT is brought up by many philosophers [3, 6, 10]. Because the TT asks for a binary result—the machine is or is not intelligent—it does not allow the interrogator to specify a level of intelligence ("this machine is as intelligent as a five-year-old"). There are two resolutions to this criticism, depending on how machine intelligence comes about. If it turns out that machines learn at a similar rate and in a similar way as humans, i.e., that there is a notion of a machine that is as intelligent as a five-year-old, but not as intelligent as an adult, then the resolution is simple: the machine should be pitted against a five-year-old in the TT. This does not change the nature of the test itself.

However, if machines turn out to learn differently from humans, or at the very least more quickly, we may not need such an intelligence gradient. For instance, a large difference between a five-year-old and an adult is their vocabulary. While it takes years for a five-year-old to develop a vocabulary comparable to an adult's, it may not for a machine. Once a machine has basic language acquisition skills and can learn to use words, it might be trivially fast to give it a large vocabulary, just as it is trivial for a machine to multiply large numbers once it has the ability to multiply at all. Once again, it is unclear whether this criticism reveals any problem with the TT, and we have no way of knowing until we build machines much more intelligent than those we have today.

## 4 The Turing Test Can Be Passed by Brute Force Machines

Next, I will deal with the criticism that the TT could be passed by an unintelligent machine. I will dub this type of machine a "brute force" machine, because it generally operates by using something akin to a large look-up table. This type of criticism usually comes from philosophers who believe it is a flaw of the TT that it only examines the machine's output. The most famous of these criticisms is Searle's Chinese Room argument [16], but Block [2] and Purtill [13] describe similar machines.

Much has been written about Searle's Chinese Room, but I believe Searle's argument–and in fact any brute-force argument—is most thoroughly debunked by Levesque in [8]. Levesque shows that the book in Searle's Chinese Room could not exist in our universe if it is in fact a look-up table. He then points out that there is a book that would work: a book that teaches Chinese. In order for machines to pass the TT, then, and interact with a human in a meaningful way through language, machines will not be able to brute-force their way through conversations; they will have to *learn* the language.

# 5 The Turing Test Is Not a Necessary Test for Intelligence

In contrast to the previous section, the next criticism falls under the umbrella of the TT being too difficult. Many philosophers (Moor [12], e.g.) believe that it is a failing of the TT that it is not a necessary condition for intelligence. I could easily dismiss this criticism by pointing out that Turing himself did not intend for the TT to be a necessary condition:

> "[The objection that the TT is not a necessary condition for intelligence] is a very strong one, but at least we can say that if, nevertheless, a machine can be constructed to play the imitation game satisfactorily, we need not be troubled by this objection."

However, dismissing the argument in this way would only leave me to address criticisms that the TT was not even a sufficient test for intelligence, all of which were dealt with in the previous section.[2] Accepting the TT as merely a sufficient test for intelligence also diminishes the test a bit, as one could simply excuse any mistakes it makes by saying "well, it is only a *sufficient* test." As a result, in subsequent sections, I will argue that the TT might in fact be a necessary test for intelligence—in the sense that it indeed tests for general intelligence, and that language captures general intelligence—but that regardless of whether it is, the TT is still worth pursuing.

# 6 How Language Skills Reflect Intelligence

In the TT, the interrogator only interacts with the machine through language; no visual or physical contact is permitted. On the one hand, this method of interaction seem fair; we should not label machines unintelligent just because they do not look like humans. On the other hand, *can* we actually separate mind and body this way? Is language expressive enough to capture all types of intelligence that we humans have?

## 6.1 The Toe-Stepping Game

One of the first philosophers to bring up the criticism that language might not capture all aspects of intelligence was Gunderson [4]. Gunderson gives an interesting analogy between the TT and a "toe-stepping game". In this game, the interrogator's toes will be stepped on by either a human or a mechanism involving a rock and a lever (which drops the rock on the interrogator's toes and then quickly removes it, as a person stepping on a toe would). The interrogator's job is to decide whether his toe is being stepped on by the human or the rock. Gunderson claims a rock could easily pass this test, and thus one might erroneously think a rock is human. He then points out that the toe-stepping game is flawed in that it has only captured one aspect of the rock's ability;

[2] The criticisms in Sections 6 and 7 deal with the fact that the TT may not test for *all* types of intelligence, but they do not claim that the TT does not sufficiently test for *some* types of intelligence.

the one aspect that is, in fact, the rock's *only* "ability". His parallel to the TT is that just as the toe-stepping game only tests one of the rock's abilities, the TT only test one of the machine's abilities: its ability to interact via language.

Gunderson's argument fails in one critical place. He is essentially making the claim that toe-stepping is analogous to linguistic capability. This is false. We know that the ability to step on toes is indicative of virtually no other ability (except, perhaps, the ability to step on other body parts), whereas linguistic capabilities are indicative of a *much* broader range of abilities [17]. One of the best ways to see this, I believe, is to note that we use language to teach other beings how to do virtually everything.[3] Another way to see this is to think closely about how we recognize intelligence in humans.

### 6.2 Recognizing Human Intelligence

One could argue that we recognize human intelligence by using the following logic: "I know that I am intelligent and I know that I am a human. Therefore, other humans are intelligent, and I recognize other humans by their appearance." Note that this test for intelligence is arguably *less* strict than the TT; we recognize human intelligence simply by looking at another being. But one might say that this logic is flawed, and that we would not deem another human intelligent until we had at least *some* interaction with them. So let us refine our argument: we *expect* other humans to be intelligent, and recognize them visually, and then upon interacting with them briefly, we *know* that they are intelligent. How do we interact with humans to know that they are intelligent? Through language. Even in criticisms of the TT, this point is acknowledged by many philosophers [12, 15, 19]. This logic works for assessing intelligence in humans because we can bootstrap the process: *I* am human, so I can ascribe intelligence to other humans.[4] To make this argument work for machines, we must argue that language captures all of human intelligence.[5]

### 6.3 What Kind of Intelligence *Doesn't* Language Capture?

At this point, there seems to be a great deal of evidence that language captures most if not all intelligence. What are some candidates for the types of intelligence that it might not capture? In [9], Michie claims that the TT doesn't capture subarticulate thought: the type of cognitive operations that we are not consciously aware of (these are similar in spirit to French's subcognitive tests in [3], but since those are used to argue that the TT only tests for human intelligence, not that language is an insufficient indicator, I will defer discussion of them to the next section). Michie argues that these types of skills are not something that we as humans can converse intelligently about: "At the very moment that a human's mastery of a skill becomes most complete, he or she becomes least capable of articulating it and hence of sustaining a convincing dialogue via the [Turing Test's] remote typewriter." Michie is in fact arguing two things in his work: that the TT cannot capture subarticulate thought—thus there is a portion of intelligence that the test misses—and, later, that machines are incapable of attaining subarticulate intelligence (see [9], pages 8–9). Note that if the latter were true, we would immediately have a way to distinguish between a human and a machine, no matter how intelligent the machine was. Through an example that Michie gives, I will show that the first argument is false, and that the second is likely false.

[3] I am not claiming that language is the most efficient way to teach certain tasks, just that it is a general tool for teaching.

[4] Modulo the Other Minds problem.

[5] In this section, I am specifically dealing with aspects of human intelligence. I will address non-human intelligence in Section 7.2.

### 6.3.1 Subarticulate Thought

Here is one example of Michie's subarticulate intelligence, from [9]: "How do you pronounce the plurals of the imaginary English words 'platch', 'snorp' and 'brell'?". The answers, obvious to any human, are "platchez", "snorpss", and "brellz". With this example, we can immediately debunk Michie's first argument. He has given us precisely a way to test for subarticulate thought: ask the machine what the plural of "platch" is.

But continuing on, is it really impossible that a machine could never achieve this type of intelligence? Michie's main defense is that we could not possibly program a machine with the correct pluralization rules. Though he in fact gives an algorithm (due to Allen, et al. [1]) for the "platch", "snorp", and "brell" case, he argues that there are "rules against which [programmers] could not have conceivably forearmed" the machine, in part because humans have not formulated all of our unconscious phonetic rules.

It is interesting that Michie uses the word "rules" here, for it does indeed seem that the way in which we pluralize words follows a set of rules, though perhaps ones that we have never explicitly written down (this hypothesis is bolstered by the fact that Michie gives us the rule for "platch"). Should that prevent a machine from acquiring the rules? It is unlikely. Given a large language corpus—perhaps the Internet—and a machine learning algorithm, it would be possible for a machine to discover countless examples of words and their plurals, and discern the phonetic rules for pluralizing. I do not mean to trivialize the task of doing so, but rather to point out that it is possible.

As an aside, it is interesting to note that building intelligent machines seems much more achievable in a world where machine learning is a commonplace technique, and where large datasets such as the Internet exist. Now it is not necessary for programmers to explicitly program in every rule about language. Instead, machines can learn from large datasets. Of course, we do not know precisely how to solve every linguistic task today, but I doubt Turing would ever have imagined that access to this type of data would be possible.

## 6.4 Turing Test Replacements

Before we move on, I'd like to briefly discuss some other human abilities that some argue would be better at assessing intelligence than language is. These are not arguments that language is a poor choice (in that it explicitly does not capture some aspect of intelligence), just that there may be a better choice.

The first is "naive psychology": the ability to ascribe intelligence to other beings. Naive psychology is generally believed to be a factor of any being possessing a mind [19], and thus any machine who is going to pass the TT should be able to exhibit skills of naive psychology. In [19], Watt actually argues that the TT as-is allows us to test for naive psychology, but that questions regarding naive psychology must explicitly be in every instance of the test, otherwise it is not a true test of intelligence. He proposes the Inverted Turing Test in [19] as an efficient way to test for naive psychology, but interestingly that test requires access to a machine that can pass today's Turing Test.

The second is a combination of language and motor skills. In [5], Harnad proposes the "Total Turing Test" (TTT), which requires a machine to possess both of these skill sets. As pointed out by Hauser in [6], Harnad actually states in his paper that linguistic capability tends to be evidence of motor skills. Thus, the TTT is redundant: if a machine could pass the TT, he could pass the TTT, assuming Harnad's conjecture is correct. In fact, Harnad goes so far to claim that body and mind are likely inseparable, so a test for only motor skills would also be redundant.

Just to address a misconception of the TT (that Harnad seems perhaps to have), nowhere in the test did Turing say that linguistic skills were the *only* skills the machine would have. If body and mind are in fact inseparable, then yes, we will not be able to create intelligent machines that have linguistic skills and not motor skills. But that is an implementation issue; it does not speak to a defect in the test.

Finally, Schweizer [15] goes one step further than Harnad, claiming that machines should have linguistic abilities, motor skills, and the ability, as a species, to create. He proposes the TTTT ("Truly Total Turing Test"), which requires an entire specifies of robots that evolve and create, and is in some sense attributing intelligence to a species, not to an individual. While that may be a fine extension of the TT, it seems difficult to generate a species of robots that could pass the TTTT without starting with a single robot that could pass the TT.

### 6.5 Summary of These Criticisms

To summarize this section, most criticisms of language as it is used in the TT claim that language does not capture all manners of intelligence. However, language is how we test for intelligence in humans, and we arguably do a less thorough job of testing when we converse with a new human than we would with a new machine. Additionally, it is not clear that language cannot capture all types of intelligence, as evidenced by my argument against Michie [9] and Hauser's [6] criticism of Harnad [5]. If, one day, we were to find a definite type of intelligence that could not be captured by language, then yes, we may want to develop a new test for intelligence. But I would argue that we must work on building intelligent machines before we can discover if this distinct type of intelligence exists (in part because it will be difficult to isolate any portion of intelligence in humans who already have linguistic abilities), and also that any new test for intelligence would likely be a stepping stone on the way to passing the TT. After all, though it would be a great achievement to pass this new intelligence test, our machine would likely be missing some key intelligences that *are* captured by the TT.

## 7 Human Intelligence vs. Other Forms

The final criticism of the TT that I will discuss is that, by nature of the interrogator being human and the machine competing against another human, the TT only captures human intelligence. That is, faced with a type of intelligence that did not resemble ours at all, the TT would fail to recognize it. But could human intelligence possibly *be* general intelligence? Would we as humans be able to recognize non-human intelligence, if it even exists?

### 7.1 The Turing Test only Tests for Human Intelligence

French claims in [3] that the TT is fundamentally a test for human intelligence. To illustrate this point, he explains the Seagull Test. In this test, the inhabitants of a Nordic island are interested in capturing the notion of flight. They know that flight is more than just remaining aloft—balloons remain aloft—or having feathers—penguins have feathers. However, their only examples of flying objects are the Nordic seagulls local to their island. They develop the Seagull Test: With two three-dimensional radar screens, they examine a seagull and a potential flying machine. If the inhabitants are unable to discern which is the seagull, then the potential flying machine can in fact fly. We can immediately see that it is highly likely that the only objects that will be able to pass this test are Nordic seagulls. As a result, the Seagull Test is really not testing for general flight; it is testing for flying Nordic seagulls. French likens this to the TT testing for human intelligence,

rather than general intelligence. I will use The Seagull Test as an example throughout the rest of this section.

### 7.2 Non-human Intelligence

Even if other forms of intelligence did exist, could we recognize them? Jacquette [7] and Watt [19] both note that we would probably not recognize non-human intelligence if not through language, as after all, that is how we recognize human intelligence (see Section 6.2). If we as humans cannot recognize non-human intelligence, we cannot expect any test we develop to do so. Similarly, though the TT is frequently generalized as a test for intelligence in anything, Turing likely meant it as a test for intelligence in *man-made* machines. Would we ever be able to create non-human intelligence in machines, without having an example of such intelligence first?

Back to the Seagull Test. How exactly did we expect the Nordic Islanders to test for flight before they had more than one example of it? One thought might be as follows: the Seagull Test is remarkably passive. Maybe if the Nordic Islanders had thought to take an active approach, that could have helped. For instance, had they observed a plane, they could have asked the pilot to perform certain maneuvers that they had seen the Nordic seagulls do. Note that the TT accomplishes this by virtue of the interrogator being an active participant. He does not simply observe the machine; he interacts with it.

Minsky brings up an interesting argument in [11]: What if human intelligence *is* actually general intelligence, or at least that any non-human intelligence could not be wildly different? He argues that this may be the case because the way humans think—dividing the world into objects—allows us to think efficiently, and that various proposals for alien intelligence—storing the world as a hologram—are remarkably *in*efficient. This is evidence that even if the TT only tests for human intelligence, that might be exactly what we want.

### 7.3 The Human Experience

In [3], French argues that not only does the TT test for human intelligence, but that one actually needs to experience the world as humans do in order to pass it. He gives many examples of "subcognitive questions" that he argues could only be answered by a human who has experienced the world. For example: "On a scale of 0 to 10, please rate 'Flugblogs' as a name Kellog's would give a new breakfast cereal" or "Rate dry leaves as a hiding place". Any human can see that "Flugblogs" is a terrible name for a cereal—it reminds us of the words "ugly" and "blob"—and that dry leaves make an excellent hiding place—we likely spent our childhood using them for that purpose.

As singular questions, I think it is unreasonable for French to believe that a machine could not answer these. As I discussed in Section 6.3.1, machines can acquire phonetic rules, making the "Flugblogs" question possible, and they can read about various human experiences from the Internet, making the dry leaves question possible. Also note that a machine will not necessarily fail the TT just because it has not had a particular human experience. We would not label a human who had never hidden in a pile of dried leaves as unintelligent, just inexperienced.

However, I will concede one point to French: If the TT interrogator asks multiple questions about the human experience to both the human and the machine, and the machine answers intelligently, but negatively, about having many of those experiences, then based solely on statistics, the interrogator will probably be able to determine which is the machine. We may be able to augment the test by simply not informing the interrogator that one of the entities he is testing is definitely a machine. If the machine is intelligent enough, but has not had many human experiences, would

the interrogator immediately think he was dealing with a machine, or just a very inexperienced human?

Regardless, it seems that the only example of intelligence we have is human intelligence as it is tied to the human experience. Let us return to the Seagull Test. What would happen if the Nordic Islanders did see a plane on their radar screen? French gives the impression that they would classify the plane as non-flying, and move on. This hardly seems likely. Though they might classify it as non-flying, one hopes that they would also classify it as something they had never seen before, and note that it was *doing* something they had never seen before (namely, remaining in the air unlike a Nordic seagull or a balloon does). The Nordic islanders would likely go out to examine the plane more closely, and hopefully after doing so, would refine their test.

This is exactly what I think would happen if the TT was presented with a different form of intelligence. We might recognize that the new intelligence did not have the same intelligence as a normal human adult; in some sense, it would fail the TT. But because of the rich way the interrogator would interact with the machine through language, he would *also* recognize it as something totally new. It seems that if we want to test for intelligence that does not require the human experience, or that is entirely unlike human intelligence, our best hope is to learn as much as possible about these types of intelligence. And one of our best hopes for learning about non-human intelligence, short of it falling from the sky, might be to try to create it. Even if it is not possible for us to create non-human intelligence (see Section 7.2), or human intelligence without human experience, we would certainly gain more understanding of it by trying.

## 8 Conclusion

My main argument in this essay has been that, at this point in time, most criticisms of the TT are wrong or irrelevant. This is admittedly a bold claim. But I believe it is supported by a few facts. First, there is much evidence that, despite criticisms, the TT is actually a very good test for intelligence. This stems from the fact that language is perhaps the best indicator we have of intelligence; one could even argue that we use less stringent tests when we identify intelligence in one another. Second, though one could rightly criticize the TT for only testing for human intelligence, it is not clear that non-human intelligence exists, nor that we as humans would even recognize it (much less that we could test for it). And third, we have no way of knowing if other criticisms are correct without having access to machines that are much more intelligent than the ones we have today, or knowing much more about intelligence.

I do not believe that an entire theory of intelligence will be necessary to pass the TT or to get to the point where we can find a definitive criticism of it. What I *do* believe is that we are nowhere near this point, and that one of the best ways to understand more about intelligence is to try to build intelligent machines, with the goal having them one day pass the TT as it is today. The knowledge that we will gain from these machines will show us how to change the TT, if needed; we cannot make effective changes to the test simply by speculating about how machine intelligence might come about.

## References

[1] J. Allen, M. S. Hunnicutt, and D. Klatt. From Text to Speech: the MITalk System. Cambridge University Press, 1987.

[2] N. Block. Psychologism and Behaviorism. In *The Philosophical Review*, volume 90, 1981.

[3] R. M. French. Subcognition and the Limits of the Turing Test. In *Mind*, volume 99, pages 53–65, 1990.

[4] K. Gunderson. The Imitation Game. In *Mind*, volume 73, pages 234–245, 1964.

[5] S. Harnad. Other Bodies, Other Minds. In *Minds and Machines*, volume 1, pages 43–54, 1991.

[6] L. Hauser. Reaping the Whirlwind. In *Minds and Machines*, volume 3, pages 219–237, 1993.

[7] D. Jacquette. Who's Afraid of the Turing Test? In *Behavior and Philosophy*, volume 20/21, pages 63–74, 1993.

[8] H. J. Levesque. Is it Enough to Get the Behaviour Right?, 1989.

[9] D. Michie. Turing's Test and Conscious Thought, 1992.

[10] P. H. Millar. On the Point of the Imitation Game. In *Mind*, volume 82, pages 595–597, 1973.

[11] M. Minsky. Communication with Alien Intelligence. In *Extraterrestrials: Science and Alien Intelligence*, 1985.

[12] J. H. Moor. An Analysis of the Turing Test. In *Philosophical Studies*, volume 30, pages 249–257, 1976.

[13] R. L. Purtill. Beating the Imitation Game. In *Mind*, volume 80, pages 290–294, 1971.

[14] A. P. Saygin, I. Cicekli, and V. Akman. Turing Test: 50 Years Later. In *Minds and Machines*, volume 10, pages 463–518, 2000.

[15] P. Schweizer. The Truly Total Turing Test. In *Minds and Machines*, volume 8, pages 263–272, 1998.

[16] J. R. Searle. Minds, Brains, and Programs. In *Behavioral and Brain Sciences*, volume 3, pages 417–457, 1980.

[17] J. G. Stevenson. On the Imitation Game. In *Philosophia*, volume 6, pages 131–133, 1976.

[18] A. Turing. Computing Machinery and Intelligence. In *Mind*, volume 59, pages 443–460, 1950.

[19] S. Watt. Naive Psychology and the Inverted Turing Test. In *Psycoloquy*, volume 7, 1996.