Does the Turing Test Demonstrate Intelligence or Not?

Stuart M. Shieber[*]
Harvard University
33 Oxford Street — 245
Cambridge, MA 02138
shieber@deas.harvard.edu

Stuart M. Shieber. 2006. Does the Turing Test Demonstrate Intelligence or Not?. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*, Boston, MA, 16–20 July.

Introduction

The Turing Test has served as a defining inspiration throughout the early history of artificial intelligence research. Its centrality arises in part because verbal behavior indistinguishable from that of humans seems like an incontrovertible criterion for intelligence, a "philosophical conversation stopper" as Dennett (1985) says. On the other hand, from the moment Turing's seminal article (Turing, 1950) was published, the conversation hasn't stopped; the appropriateness of the Test has been continually questioned, and current philosophical wisdom holds that the Turing Test is hopelessly flawed as a sufficient condition for attributing intelligence.

In this short article, I summarize for an artificial intelligence audience an argument that I have presented at length for a philosophical audience (Shieber, to appear) that attempts to reconcile these two mutually contradictory but well-founded attitudes towards the Turing Test that have been under constant debate since 1950 (Shieber, 2004).

The arguments against the sufficiency of the Turing Test for determining intelligence rely on showing that some extra conditions are logically necessary for intelligence beyond the behavioral properties exhibited by a subject under a Turing Test. Therefore, it cannot follow logically from passing a Turing Test that the agent is intelligent. I will argue that these extra conditions *can* be revealed by the Turing Test, so long as we allow a very slight weakening of the criterion from one of logical proof to one of statistical proof under weak realizability assumptions. Crucially, this weakening is so slight as to make no conceivable difference from a practical standpoint. Thus, the Gordian knot between the two opposing views of the sufficiency of the Turing Test can be cut.

The Essence of the Turing Test

The Turing Test is, at its heart, a test of the adequacy of an agent's verbal behavior. Block (1981) characterizes it as a test of the ability to "produce a sensible sequence of verbal responses to a sequence of verbal stimuli". Turing's own descriptions of the Test (Newman *et al.*, 1952) accord with this view:

> The idea of the test is that the machine has to pretend to be a man, by answering questions put to it, and it will only pass if the pretence is reasonably convincing.... We had better suppose that each jury has to judge quite a number of times, and that sometimes they really are dealing with a man and not a machine. That will prevent them saying "It must be a machine" every time without proper consideration.

Turing's original presentation of the test is couched in terms of an imitation game between two entities, a person and a machine, with the goal of seeing if in repeated forced choices a judge can do no better than chance at determining which is which on the basis of verbal interactions with each. The introduction of the human confederate and the forced choice merely serve to make more clear and operational what constitutes "sensibility" of the machine's responses.

Thus, at base, the Turing Test is founded on the idea that ability to produce sensible verbal behavior is an indication of intelligence. The syllogism that underlies the appropriateness of the Turing Test as a criterion for intelligence proceeds something like this:

Premise 1: If an agent passes a Turing Test, then it produces a sensible sequence of verbal responses to a sequence of verbal stimuli.

Premise 2: If an agent produces a sensible sequence of verbal responses to a sequence of verbal stimuli, then it is intelligent.

Conclusion: Therefore, if an agent passes a Turing Test, then it is intelligent.

Block refers to a premise such as the second one as the "Turing Test conception of intelligence", and his (and others') repudiation of the Turing Test as a criterion for intelligence is based on a denial of this premise.

For instance, passing a Turing Test on a single occasion might be the result of chance. Even monkeys on typewriters might "produce a sensible sequence of verbal responses" on (astronomically rare) occasion. (For this reason, Turing would have multiple tests be run.) Thus, the premise ought to be interpreted as a *general capacity*, not an occasional feat. But as we will see, even this reinterpretation of Premise 2 is insufficient.

[*] I gratefully acknowledge by reference the large number of people acknowledged in my paper on which this summary is based (Shieber, to appear).

The Argument Against a Behaviorist Test

The anti-behaviorist argument against the Turing Test as a sufficient condition for intelligence was apparently first proposed in sketch form by Shannon & McCarthy (1956, page vi): "A disadvantage of the Turing definition of thinking is that it is possible, in principle, to design a machine with a complete set of arbitrarily chosen responses to all possible input stimuli.... With a suitable dictionary such a machine would surely satisfy Turing's definition but does not reflect our usual intuitive concept of thinking."

In "Psychologism and Behaviorism", Block (1981) presents the argument in its most fully worked out form. Imagine (with Block) a hypothetical machine that stores a tree of interactions providing a sensible response for each possible interrogator's input in each possible conversational context of up to, say, one hour long. (These responses might be modeled on those that Block's fictional Aunt Bertha would have given.) Such a tree would undeniably be large, but processing in it would be conceptually straightforward. By hypothesis, such an "Aunt Bertha machine" would pass a Turing Test of up to one hour, because its responses would be indistinguishable from that of Aunt Bertha, whose responses it recorded. Such a machine is clearly not intelligent—Block (1981) says it "has the intelligence of a toaster", and Shannon and McCarthy would presumably agree—by the same token that the teletype that the interrogator interacts with in conversation with the human confederate in a Turing Test is not intelligent; it is merely the conduit for some other person's intelligence, the human confederate. Similarly, the Aunt Bertha machine is merely the conduit for the intelligence of Aunt Bertha. Yet just as surely, it can pass a Turing Test, and more, has the *capacity* to pass arbitrary Turing Tests of up to an hour. The mere logical possibility of an Aunt Bertha machine is sufficient to undermine premise 2, even under a reinterpretation as requiring a general capacity.

It seems to me that Shannon, McCarthy, and Block are right in principle: Such a machine is conceptually possible; hence the Turing Test is not *logically* sufficient as a condition of intelligence. Let us suppose this view is correct and, as Block argues, some further criterion is needed regarding the manner in which the machine works. Some further criterion is needed, but how much of a criterion is that, and can the Turing Test test for it? Although Block calls this further internal property "nonbehavioral", I will argue that *the mere behavior of passing a Turing Test can reveal the property*. Borrowing an idea from theoretical computer science, I argue that the Turing Test can be viewed as an *interactive proof* not only of the fact of sensible verbal behavior, but of a capacity to generate sensible verbal behavior, and to do so "in the right way". Assuming some extraordinarily weak conditions on physical realizability, any Turing-Test–passing agent must possess a sufficient property to vitiate Block's argument. In summary, Block's arguments are not sufficient to negate the Turing Test as a criterion of intelligence, at least under a very slight weakening of the notion of "criterion".

Block pursues a number of potential objections to his argument, the most significant of which (his "Objection 8") is based on the fact that the Aunt Bertha machine is exponentially large, that is, its size is exponential in the length of the conversation. Objection 8 leads to his "amended neo-Turing-Test conception": "Intelligence is the capacity to emit sensible sequences of responses to stimuli, *so long as this is accomplished in a way that averts exponential explosion of search*." (Emphasis in original.) It is not exactly clear what "exponential explosion of search" is intended to indicate in general. In the case of the Aunt Bertha machine, exponentiality surfaces in the size of the machine, not the time complexity of the search. Further, the aspect of the Aunt Bertha machine that conflicts with our intuitions about intelligence is its reliance upon *memorization*. Removing the possibility of exponential storage amounts to a prohibition against memorization.[1] Consequently, an appropriate rephrasing of Premise 2 is

The compact conception of intelligence: If an agent has the capacity to produce a sensible sequence of verbal responses to a sequence of verbal stimuli, whatever they may be, and without requiring storage exponential in the length of the sequence, then the agent is intelligent.

Again, Block notes that this additional condition is psychologistic in mentioning a nonbehavioral condition, viz., that the *manner* of the processing must avert combinatorial explosion of storage. He claims that insofar as the condition is psychologistic, a Turing Test cannot test for it.

To summarize, Block's Aunt Bertha argument forces us to pay up on two psychologistic promissory notes. For the purely behavioral Turing Test to demonstrate intelligence, it must suffice as a demonstration of the antecedent of the compact conception of intelligence, that is, it must indicate a *general capacity* to produce a sensible sequence of verbal responses and it must demonstrate *compactness* of storage of the agent.

The Interactive Proof Alternative

Turning to the capacity issue first, there is certainly no deductive move that allows one to go from observation of the passing of one or more Turing Tests to a conclusion of a general capacity; the monkeys and typewriters argument shows that. This is the Humean problem of induction. But it does not follow that there is no method of reasoning from the former to the latter. I will argue that the powerful notion of an *interactive proof*, taken from theoretical computer science, is exactly such a reasoning method. Furthermore, as discussed below, Turing Tests bear some of the tell-tale signs of interactive proofs that have been investigated in the computer science literature.

Interactive proofs are protocols designed to convince a verifier conventionally denoted V that a prover P has certain knowledge or abilities, which we will think of as being encapsulated in an assertion s.[2] In a classical (deductive) proof

[1] For this reason, adding this extra condition to the conception of intelligence is not ad hoc. It amounts to saying, in a precise way, that the agent must have the capacity to produce sensible responses without having memorized them.

[2] For convenience in reference, we will refer to V and P using gendered pronouns "she" and "he" respectively.

system, P would merely reveal a deductive proof of s, which V then verifies. This provides V with knowledge of s and perhaps other knowledge implicit in the proof. Interactive proofs augment classical proof systems by adding notions of *randomization* and *interaction* between prover and verifier. (The interaction implicit in classical proof systems — P's presenting V with the proof — is essentially trivial.) Interaction is added by allowing V and P to engage in rounds of message-passing. Randomization is introduced in two ways: First, the verifier may make use of random bits in constructing her messages. Second, she may be required to be satisfied with a probabilistic notion of proof. When we state that P proves s with an interactive proof, we mean (implicitly) that s has been proved but with a certain determinable residual probability of error. That is, the verifier may need to be satisfied with some small and quantifiable chance that the protocol indicates that s is true when in fact it is not, or vice versa. The residual error is the reason that moving to a notion of interactive proofs is a weakening relative to a view as a deductive proof. The fact that the residual error can rapidly be made vanishingly small through repeated protocols is the reason that the weakening is referred to as "very slight".

The Turing Test as an Interactive Proof of Capacity

I view the Turing Test as an interactive proof for the antecedent of the capacity conception of intelligence, that is, it is a proof that P "has the capacity to produce a sensible sequence of verbal responses to a sequence of verbal stimuli, whatever they may be". Consider the space of all possible sequences of verbal stimuli. Let the fraction of this space for which P generates sensible responses be t_p. An agent without a general capacity to produce sensible sequences of responses would fail to do so on some nontrivial fraction of this space. Block notes that a 100% criterion is neither necessary nor appropriate. One wants to be able to "ask of a system that fails the test whether the failure *really does* indicate that the system lacks the disposition to pass the test." Indeed, people put under similar tests would at least occasionally perform in such a way that a judge might deem the responses not sensible. So there is some percentage t_l, less than 100%, such that if an agent produced sensible sequences of responses on that percentage of the space (that is, $t_p > t_l$), we can attribute a general capacity, sufficient for the antecedent of the capacity conception. Let us say, for the sake of argument, that $t_l = 1/2$. Thus, if an agent produces sensible responses to 50% of the space of possible verbal stimuli, we will consider it to have a general capacity to produce such responses. Importantly, we are not saying that the agent must merely produce sensible responses to 50% of some subsample of possible stimuli that we confront it with, but with 50% of all possible stimuli, in a counterfactual sense, whether we ever test it with these stimuli or not.

Suppose we sample k sequences of verbal stimuli uniformly from this space, and test some agent as to whether it generates sensible sequences of responses to them. Suppose further that the agent does so on t of these stimuli, where t is greater than a sample threshold $t_s > t_l$ (say, $t_s = 3/4$). Can we conclude that the agent has a general capacity as defined above? A false positive occurs when a sample of k inputs is selected where $t > t_s$ (the prover outperforms the sample threshold on the sample), yet $t_p < t_l$ (the subspace is smaller than the definitional threshold). Using the method of Chernoff bounds (see, e.g., Chapter 5 of the text by Motwani (1995)), it can be shown that the probability of a false positive is $Pr[t > t_s] < e^{-ck}$ where $c = \frac{(t_l - t_s)^2}{2(1-t_l)}$. Thus, it has the behavior of an interactive proof: As the number of samples k increases, the probability of a false positive decreases exponentially.

It is important to realize that the probabilities of error that we are talking about can be literally astronomically small. For the bounds that we have been talking about, if we let k be, say, 300, the false positive probability is on the order of 1 in 10^{10}; at that rate, if a population the size of all humanity were tested, one would expect to see *no* false positives. At $k = 1000$, the false positive rate of some 1 in 10^{27} is truly astronomically small.

In summary, a protocol in which we run k Turing Tests and receive sensible responses on an appropriate fraction provides exponentially strong evidence that the agent satisfies the antecedent of the capacity conception, that is, has a general capacity to produce sensible responses to verbal stimuli, whatever they may be. One may quibble with the various bounds proposed. (Should t_l be 50%? 80%? 99%?) Varying them does not change the basic character of the argument. It merely adjusts the number of samples needed before the knee in the exponential curve.

Thus, under the notion of proof provided by interactive proofs, the Turing Test can provide a proof of a *general capacity* to produce a sensible sequence of verbal responses to a sequence of verbal stimuli, whatever they may be. It can therefore unmask the monkeys on typewriters.

The Turing Test as an Interactive Proof of Compactness

The interactive proof approach provides leverage for demonstrating compactness as well. When all we know is the agent's performance on a fixed sample of stimuli, the storage requirements to generate these responses is linear in the length of the stimuli. But the size of any fixed *fraction* of the space of possible stimuli is exponential in their length. By being able to reason from the sample to the fraction of the space as a whole — as the interactive proof approach allows — we can conclude that an agent using a memorization strategy (as the Aunt Bertha machine) would require exponential storage capacity to achieve this performance. Conversely, any agent not possessing exponential storage capacity would fail the interactive proof.

Nonetheless, how can a Turing Test reveal that the machine *doesn't* have exponential storage capacity? The compact conception would require that the agent pass Turing tests *at least logarithmic in its storage capacity*. Thus, without bounding its storage capacity, we can't bound the length of the Test we would need. There is no purely logical argument against this possibility; the Aunt Bertha argument shows this. Some further assumption must be made to pay the compactness promissory note. I now turn to how weak an assumption is required.

Suppose we could bound the information capacity of the universe. Then any physically realizable agent that could pass Turing Tests whose length exceeded the logarithm of this amount would satisfy the compact conception. We would be able to bound the length of the Turing Test required under the compact conception, at least for any agent that is *no larger than the universe*. (And of course, no agent is larger than the universe.) We will call this length bound the *critical Turing Test length*. One might worry that the critical Turing Test length might be centuries or millennia long.

Without going into detail (which I have provided elsewhere (Shieber, to appear)), the information capacity of the universe can be estimated based on the holographic principle (regarding which see the survey by Bousso (2002) for a review) and estimates of the time since the Big Bang. Together, these lead to a reasonable upper bound of some 10^{120} bits. Rounding up by 80 orders of magnitude, call it 10^{200}.

Descending now from these ethereal considerations to the concrete goal of analyzing the Turing Test conceptions of intelligence, under the compact conception, we would require an agent with this literally astronomical storage capacity to have a capacity to pass Turing Tests of on the order of $\log_2 10^{200} \approx 670$ bits. The entropy of English is about one bit per character or five bits per word (Shannon, 1951), so we require a critical Turing Test length of around 670 characters or 140 words. At a natural speaking rate of some 200 words per minute, a conversation of less than a minute would therefore unmask a Turing-Test subject whose performance, like that of the Aunt Bertha machine, is based on memorization.

In essence, I have argued for the following recasting of the basic syllogism supporting the sufficiency of the Turing Test:

Premise 1: If an agent passes *k* rounds of a Turing Test of at least one minute in length, then (with probability of error exponentially small in *k*) it has the capacity to produce a sensible sequence of verbal responses to a sequence of verbal stimuli that is logarithmic in the storage capacity of the agent, whatever they may be.

Premise 2: If an agent has the capacity to produce a sensible sequence of verbal responses to a sequence of verbal stimuli that is logarithmic in the storage capacity of the agent, whatever they may be, then it is intelligent.

Conclusion: Therefore, if an agent passes *k* rounds of a Turing Test of at least one minute in length, then (with probability of error exponentially small in *k*) it is intelligent.

As contributory evidence for this view of the Turing Test as an interactive proof, I note that the Turing Test shares various other properties with interactive proofs, such as

Nontransferability: Turing tests, like other interactive proofs, provide proof only to the verifier, and not third parties. This property is familiar to those of us in the natural-language-processing field as the "cooked demo".

Lack of closure under composition: Turing Tests, like other interactive proofs, may lose their proof characteristics under composition. For example, Turing Tests are subject to a "man-in-the-middle attack", by which an agent that can't pass a singleton Turing Test can do so by simultaneously participating as judge in another.

Conclusion

I have argued that the Turing Test is appropriately viewed not as a deductive or inductive proof but as an interactive proof of the intelligence of a subject-under-test. This view is evidenced both by the similarity in form between Turing Tests and interactive proof protocols and by the sharing of important properties between Turing Tests and interactive proofs.

In so doing, I provide a counterargument against Block's demonstration that the Turing Test is not a sufficient criterion of intelligence. The counterargument requires a (very slight) weakening of the conditions required of the Turing Test — weakening the notion of proof (from classical deductive proof to interactive proof with its exponentially small residual error probability) and strengthening the notion of possible agent (from one of logical possibility to one with a trivial realizability requirement). These weakenings are sufficiently mild that they can be seen as providing foundation for the view that the Turing Test is a sound sufficient condition for intelligence. Block is right, yet Dennett may be too.

It merits pointing out that this view of the Turing Test is consonant with (though by no means implicit in) Turing's view of the Test as presented in his writings. His view of the Test as being statistical in nature and his pragmatic orientation toward its efficacy are of a piece with its status as an interactive rather than classical proof.

References

Block, N. 1981. Psychologism and behaviorism. *Philosophical Review* XC(1):5–43.

Bousso, R. 2002. The holographic principle. *Reviews of Modern Physics* 74:825–874. Available as hep-th/0203101.

Dennett, D. 1985. Can machines think? In Shafto, M., ed., *How We Know*. San Francisco, CA: Harper & Row. 121–145.

Motwani, R. 1995. *Randomized Algorithms*. Cambridge, England: Cambridge University Press.

Newman, M. H. A.; Turing, A. M.; Jefferson, S. G.; and Braithwaite, R. B. 1952. Can automatic calculating machines be said to think? Radio interview, recorded 10 January 1952 and broadcast 14 and 23 January 1952. Turing Archives reference number B.6.

Shannon, C. E., and McCarthy, J., eds. 1956. *Automata Studies*. Princeton, NJ: Princeton University Press.

Shannon, C. E. 1951. Prediction and entropy of printed English. *Bell Systems Technical Journal* 30(1):50–64.

Shieber, S. M. 2004. *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*. MIT Press.

Shieber, S. M. To appear. The Turing test as interactive proof. *Noûs*.

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