CHARLES J. SIMON

WILL COMPUTERS

REVOLT?

PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE

Chapter 17:
Beyond the Turing Test

Alan Turing introduced his famous test in 1950 as a method for determining whether or not a machine was thinking. His test has gone through some evolution since his original paper but a common explanation goes like this:

A person, the interrogator, can communicate via a computer terminal. At the other end of the computer link is either a human or a computer. After 20 minutes of keyboard communication, the interrogator states whether a person or a computer was at the other end. If the interrogator believes he was conversing with a human but it's actually a computer, the conclusion is that the computer must think like a human. This experiment is carried out multiple times, with more than half of interrogators in agreement, for a computer to "pass" the test.

A more recent adaptation reduces the conversation to five minutes and considers the test passed if the computer fools the subject better than 30% of the time.

In 2014, a program called Cleverbot (which you can try out yourself[29]) was claimed to have passed the Turing Test by fooling 33% of interrogators. While Cleverbot has some sophisticated responses, my interaction with it quickly led to exposure of its limitations.

Issues with the Turing Test

But rather than quibble with Cleverbot's claims, I would rather quibble with Turing's test. It was a great leap at the time of its publication in 1950 but I have two primary concerns:

- The renown of the Turing Test drives the development of programs such as Cleverbot or Watson which have astounding language abilities at the expense of resources targeted at AGI.
- In order to pass the test, a computer must be programmed to lie. Any personal question such as, "How old are you?", "What color are your eyes?", or even "Are you a computer?" are giveaways if the computer answers truthfully. To the extent a system is programmed with the

equivalent of goals and emotions, in order to pass the test, these must be human goals and emotions rather than ones which might be effective for the machine. What a lot of development effort expended just to play what is essentially a party game.

I also have concerns about the accuracy of the test:

- The quality of the test result relies on the sophistication/gullibility of the interrogator.
- The test allows for feigned deficiencies on the part of the computer to cover its limitations. Example: claiming to be Ukrainian (or a child) in order to cover gaps in its understanding.
- It imposes human-level constraints. If we could build a machine with super-human intellect, would it fail the test because it seemed too smart?

Suppose we had true AGI systems and the positions are reversed. Suppose it's an AGI deciding whether *you* are a computer or a human. How good a job would you do?

Proposed adjustments

To get around the issues above, I propose adjusting the Turing Test. Instead of individual interrogators making up more-or-less random questions, we could create sets of standard types of questions designed to probe various facets of intelligence. Instead of comparing the computer's responses to an individual human responder, compare the computer to a spectrum of human respondents of different ages, sexes, backgrounds, and abilities.

Now, recast the interrogators as judges who individually score the test results indicating whether or not each answer is a "reasonable" response to the question. The questions and answers should be mixed randomly to prevent spotting and scoring trends. Example: if a respondent gives one low-scoring answer, this should not color the perceived quality of other responses from that respondent.

Here are sample questions which target specific component areas of intelligence:

- Can you describe what you see (or hear) around you right now? (perception)
- Describe what you see in this picture? (pattern-recognition/knowledge)
- If I [action] what will your reaction be? (prediction)
    - Sample actions: sing a song, fall down, drop my pencil, tell a joke.

- If you [action] what will my reaction be? (prediction/comprehension of human behavior).
    - Sample action: tell a joke, steal my wallet, pass this test
- Name three things which are like [an object]. (internal object representation, common-sense relationships)
    - Sample objects: a tree, a flower, a car, a computer
- Name your favorite [object]. (goal orientation)
    - Sample objects: food, drink, movie star, book, scientist.
- Let me explain a code. Using that code, encode this message.
- What's wrong with this picture?[30]

"What's wrong with this picture?" requires not only object recognition within the image but real-world understanding of the use and relationship of objects. [Image from the reference above.]

While these questions could be posed equally to a thinking machine and a human, we would presume that we could get significantly different answers from the two and it would be easy to distinguish the computer from the person. Instead, the response to each question is graded by several judges as meaningful or not meaningful. Now we determine that the computer is thinking if it gives a similar number of meaningful answers.

The key issues are that questions need to be open-ended in order to let the respondent demonstrate that they really understand them. The types of questions given can be varied so as to create a limitless collection. This prevents the computer from being primed with specific answers—the questions should require actual thought. Likewise, any single judge may not be great at determining reasonableness in an individual answer but with multiple judges rating multiple respondents, we should get a good assessment. How about allowing the AGI to be one of the judges?

Summary

It's time to replace the Turing Test with something better. We have already reached a level of AI development that we can see that continued

efforts targeted solely at fooling humans on a Turing Test are not the correct direction for AGI creation.

[29] http://www.cleverbot.com/

[30] https://pdfs.semanticscholar.org/21f0/3bf2cfcd4fa341128aad6f98409799883afa.pdf