



A student led computer science society dedicated to exploring computing, technology, and the ideas shaping our digital world.

Von Neumann Society - CCGS

Is the Turing test still a meaningful measure of intelligence?

Research Report by Rex Beevers

In 1950, the British computer scientist Alan Turing asked the question: Can machines think? Instead of tackling the philosophical problem of defining “thinking”, he came up with a simple test. A human judge has a conversation with two hidden partners, one machine and one human. If the judge cannot tell which is which, the machine passes the test.

For a long time, the Turing test was seen as a good measure for machine intelligence, but in today’s age of AI chatbots and LLMs, is the test still a meaningful measure of intelligence? With chatbots such as ChatGPT being able to hold long lasting, human-like conversations, does the test still tell us anything useful about intelligence, or just how good a machine is at pretending to be human?

Why Turing’s idea mattered

In 1950, computers were extremely basic, and the meaning of “thinking” was not even agreed upon. Even today the question of what thinking actually is, is still an ongoing philosophical debate. This is why Turing’s move to change the question from “Can machines think?”, to “Can machines behave as if they think?” was so significant. He turned the problem from an impossible philosophical question to something you could test with an experiment.

Turing deliberately did not define thinking. He pointed out that we cannot directly see another person's thoughts, we can only infer their consciousness from how they behave. So, he argued, why should a machine be treated differently?



A student led computer science society dedicated to exploring computing, technology, and the ideas shaping our digital world.

Von Neumann Society - CCGS

Arguments for the Turing test

Today there is still not an agreed upon definition of thinking or consciousness. The Turing test cleverly avoids philosophy entirely by judging the behaviour of the machine instead, and giving us an experiment we can measure.

Turing extended this same logic to machines. If a human is judged as intelligent purely by displaying behaviour we associate with thinking such as, language, reasoning, problem solving, humour, and a machine can convincingly display this behaviour and pass the test, it could be considered good evidence of intelligence.

It complements other tests, rather than replacing them. The Turing test focuses mainly on the language of a machine. Language is only one component of intelligence, and the Turing test is a very valid way to judge a machine's language. It can be coupled with other tests for mathematics and problem solving to give a more meaningful measure of machine intelligence.

ELIZA: Imitation without understanding

One of the earliest indications fooling a judge is not the same as being intelligent came from the 1966 chatbot ELIZA. Created by Joseph Weizenbaum, ELIZA simulated a Rogerian psychotherapist and used simple pattern-matching to reply to the user. It had no real understanding of the user's messages, yet some users treated its replies as if it did. Weidenbaum's secretary even asked him to leave the room to speak to ELIZA privately.

ELIZA's replies looked like the responses of something that understood, but it clearly did not. This raised an important question:

"Can a machine imitate a human without possessing any real intelligence?"

The Chinese Room: Imitation is not understanding

In 1980, philosopher John Searle designed a thought experiment, called the Chinese Room, to attack this exact question.

Imagine you are locked in a room. You do not speak any Chinese. Notes written in Chinese are slipped through a slot in the room. You have a giant rule book that tells you exactly what



A student led computer science society dedicated to exploring computing, technology, and the ideas shaping our digital world.

Von Neumann Society - CCGS

symbols to write back. With this rulebook, you can produce perfect responses in Chinese, and to a person outside the room you seem to understand Chinese fluently. However, you have no understanding of Chinese, you were only following the rulebook.

This is essentially how ELIZA and to an extent all computers work. A machine can follow rules (its program) to produce the right output without ever understanding the meaning of the input. This directly challenges Turing's test.

Turing's view

If a machine behaves intelligently, that may be evidence to call it intelligent.

Searle's view

A machine can behave intelligently without understanding anything at all, so behaviour alone is not sufficient evidence to prove intelligence.

Problems with the Turing test

The test measures imitation not intelligence. Fooling a human judge shows a machine is good at pretending to be human, but not necessarily that the machine is smart. Additionally, it focuses too much on language. An AI could excel at maths or problem solving and still fail the Turing test if its language is lacking. The test focuses too much on one part of human behaviour to give an accurate representation of machine intelligence.

The judge makes a difference. 2 people could be given the exact same conversations and choose different outcomes, making the test subjective.

How this links to modern AI

This argument is more relevant today than ever as the field of AI is constantly changing. Modern chatbots called large language models (LLMs), such as ChatGPT and Gemini, do not think the way humans do. They are trained on huge amounts of human-written text and data, to predict what word is most likely to come next in a sentence. This lets them produce extremely fluent, human-like answers.



A student led computer science society dedicated to exploring computing, technology, and the ideas shaping our digital world.

Von Neumann Society - CCGS

The theory behind LLMs is very similar to the Chinese Room. Instead of a rulebook, LLMs use statistical patterns but the basic idea is the same. They can produce highly convincing language without understanding what any of it means.

A 2025 study found that ChatGPT-4.5 was mistaken for human more often than real humans were. This shows passing the Turing test has become easier for AI, without settling the question of whether these machines truly think or understand at all.

Conclusion

The Turing test was a brilliant idea for its time, turning an impossible philosophical question into a measurable test. But the Chinese Room argument shows a serious flaw. A machine can imitate human behaviour without any understanding. Today's LLMs make this problem very real, as they can imitate intelligent conversation perfectly, without necessarily thinking. For these reasons, the Turing test alone is no longer enough to measure machine intelligence. However, it has been a useful tool for showing how machines have changed over time, as the percentage of LLMs being able to pass this test has drastically increased since ELIZA in 1966.