How Smart Are the Robots Getting?

The Turing test used to be the gold standard for proving machine intelligence. This generation of bots is racing past it.

-

By Cade Metz Jan. 20, 2023 NYT

Franz Broseph seemed like any other Diplomacy player to Claes de Graaff. The handle was a joke — the Austrian emperor Franz Joseph I reborn as an online bro — but that was the kind of humor that people who play Diplomacy tend to enjoy. The game is a classic, beloved by the likes of John F. Kennedy and Henry Kissinger, combining military strategy with political intrigue as it recreates the First World War: Players negotiate with allies, enemies and everyone in between as they plan how their armies will move across 20th-century Europe.

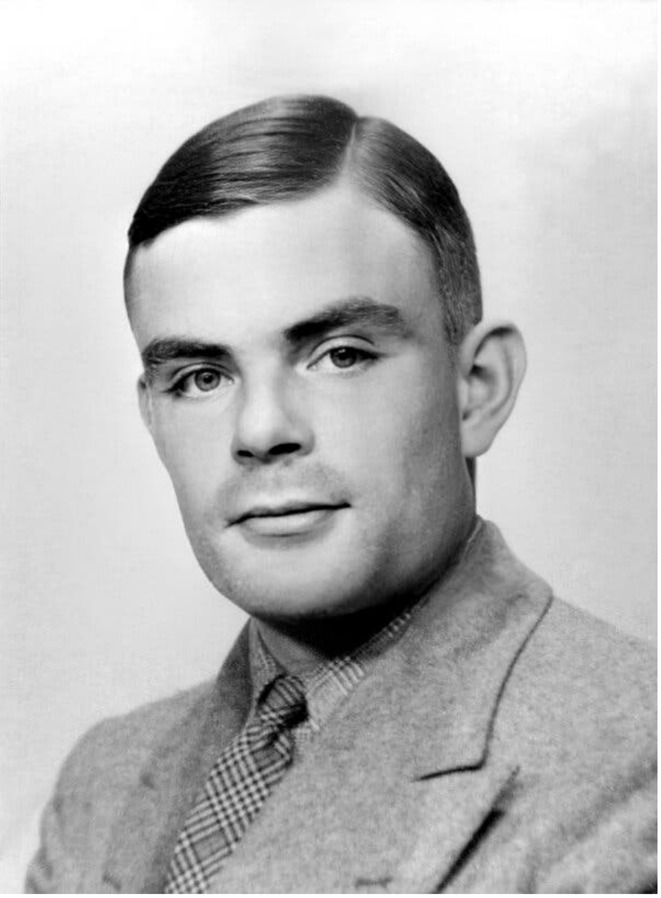
When Franz Broseph joined a 20-player online tournament at the end of August, he wooed other players, lying to them and ultimately betraying them. He finished in first place.

Mr. de Graaff, a chemist living in the Netherlands, finished fifth. He had spent nearly 10 years playing Diplomacy, both online and at face-to-face tournaments across the globe. He did not realize until it was revealed several weeks later that he had lost to a machine. Franz Broseph was a bot.

“I was flabbergasted,” Mr. de Graaff, 36, said. “It seemed so genuine — so lifelike. It could read my texts and converse with me and make plans that were mutually beneficial — that would allow both of us to get ahead. It also lied to me and betrayed me, like top players frequently do.”

Built by a team of artificial intelligence researchers from the tech giant Meta, the Massachusetts Institute of Technology and other prominent universities, Franz Broseph is among the new wave of online chatbots that are rapidly moving machines into new territory.

When you chat with these bots, it can feel like chatting with another person. It can feel, in other words, like machines have passed a test that was supposed to prove their intelligence.



Alan Turing, a British mathematician, proposed in 1950 that the test of machine intelligence would be an ability to conduct a conversation in an indistinguishably human way.Credit...Archivio GBB, via Alamy

For more than 70 years, computer scientists have struggled to build technology that could pass the Turing test: the technological inflection point where we humans are no longer sure whether we are chatting with a machine or a person. The test is named for Alan Turing, the famed British mathematician, philosopher and wartime code breaker who proposed the test back in 1950. He believed it could show the world when machines had finally reached true intelligence.

The Turing test is a subjective measure. It depends on whether the people asking the questions feel convinced that they are talking to another person when in fact they are talking to a device.

But whoever is asking the questions, machines will soon leave this test in the rearview mirror.

Bots like Franz Broseph have already passed the test in particular situations, like negotiating Diplomacy moves or calling a restaurant for dinner reservations. ChatGPT, a bot released in November by OpenAI, a San Francisco lab, leaves people feeling as if they were chatting with another person, not a bot. The lab said more than a million people had used it. Because ChatGPT can write just about anything, including term papers, universities are worried it will make a mockery of class work. When some people talk to these bots, they even describe them as sentient or conscious, believing that machines have somehow developed an awareness of the world around them.

Privately, OpenAI has built a system, GPT-4, that is even more powerful than ChatGPT. It may even generate images as well as words.
And yet these bots are not sentient. They are not conscious. They are not intelligent — at least not in the way that humans are intelligent. Even people building the technology acknowledge this point.

These bots are pretty good at certain kinds of conversation, but they cannot respond to the unexpected as well as most humans can. They sometimes spew nonsense and cannot correct their own mistakes. Although they can match or even exceed human performance in some ways, they cannot in others. Like similar systems that came before, they tend to complement skilled workers rather than replace them.

Part of the problem is that when a bot mimics conversation, it can seem smarter than it really is. When we see a flash of humanlike behavior in a pet or a machine, we tend to assume it behaves like us in other ways, too — even when it does not. The Turing test does not consider that we humans are gullible by nature, that words can so easily mislead us into believing something that is not true.

The Rise of OpenAI

The San Francisco company is one of the world's most ambitious artificial intelligence labs. Here's a look at some recent developments.

- **ChatGPT:** The cutting-edge chatbot is raising fears of students cheating on their homework. But its potential as an educational tool outweighs its risks, our technology columnist writes.
- **DALL-E 2:** The system lets you create digital images simply by describing what you want to see. But for some, image generators are worrisome.
- **GPT-3:** With mind-boggling fluency, the natural-language system can write, argue and code. The implications for the future could be profound.

"These systems can do a lot of useful things," said Ilya Sutskever, chief scientist at OpenAI and one of the most important A.I. researchers of the past decade, referring to the new wave of chatbots. "On the other hand, they are not there yet. People think they can do things they cannot."

As the latest technologies emerge from research labs, it is now obvious — if it was not obvious before — that scientists must rethink and reshape how they track the progress of artificial intelligence. The Turing test is not up to the task.

Time and time again, A.I. technologies have surpassed supposedly insurmountable tests, including mastery of chess (1997), "Jeopardy!" (2011), Go (2016) and poker (2019). Now it is surpassing another, and again this does not necessarily mean what we thought it would.

We — the public — need a new framework for understanding what A.I. can do, what it cannot, what it will do in the future and how it will change our lives, for better or for worse.

The Imitation Game

The 2014 film “The Imitation Game,” which depicts Turing’s successful attempt to crack the German Enigma code during World War II, starred Benedict Cumberbatch.Credit...Jack English/Weinstein Company

In 1950, Alan Turing published a paper called “Computing Machinery and Intelligence.” Fifteen years after his ideas helped spawn the world’s first computers, he proposed a way of determining whether these new machines could think. At the time, the scientific world was struggling to understand what a computer was. Was it a digital brain? Or was it something else? Turing offered a way of answering this question.

He called it the “imitation game.”

It involved two lengthy conversations — one with a machine and another with a human being. Both conversations would be conducted via text chat, so that the person on the other end would not immediately know which one he or she was talking to. If the person could not tell the difference between the two as the conversations progressed, then you could rightly say the machine could think. “The question and answer method seems to be suitable for introducing almost any one of the fields of human endeavour that we wish to include,” Turing wrote. The test could include everything from poetry to mathematics, he explained, laying out a hypothetical conversation:

Q: Please write me a sonnet on the subject of the Forth Bridge.

A: Count me out on this one. I never could write poetry.

Q: Add 34957 to 70764.
A: (Pause about 30 seconds and then give as answer) 105621.

Q: Do you play chess?
A: Yes.

Q: I have K at my K1, and no other pieces. You have only K at K6 and R at R1. It is your move. What do you play?
A: (After a pause of 15 seconds) R-R8 mate.

When Turing proposed the test, computers could not chat. Scientists communicated with these room-size machines by feeding mathematical and textual instructions into vacuum tubes via typewriters, magnetic tape and punched cards. But as the years passed and researchers created a new field they called artificial intelligence — a concerted effort to build machines that could think at the level of a human — many held up the imitation game as the ultimate goal.

“People were not building systems for fluent dialogue. That was just too hard,” said Stuart Shieber, a Harvard computer scientist who specializes in computational linguistics, including the Turing test. “But it was an aspiration.”

By the mid-1960s, machines could chat in small ways. And even then, they fooled people into believing they were more intelligent than they really were.

Joseph Weizenbaum, a professor at M.I.T., invented the natural language program Eliza in the 1960s.Credit...Calvin Campbell/Massachusetts Institute of Tech.

A researcher at the Massachusetts Institute of Technology, Joseph Weizenbaum, built an automated psychotherapist called Eliza, which did little more than repeat what a user said in the form of a question. But some treated this bot as if it were a human therapist, unloading their most personal secrets and feelings.

Over the next several decades, chatbots improved at a snail's pace. The best that researchers could do was lay down a long list of rules defining how a bot should behave. And no matter how many rules they wrote, they were never enough. The scope of natural language was just too big.

In 2014, after nearly 60 years of A.I. research, three researchers in St. Petersburg, Russia, built a bot, called Eugene Goostman, that imitated a 13-year-old Ukrainian who had learned English as a second language. But claims from its creators — and from the news media — that it had passed the Turing test were greatly exaggerated. When asked, "Which is bigger, a shoe box or Mount Everest?," this bot said: "I can't make a choice right now." When asked, "How many legs does a camel have?," it replied: "Something between 2 and 4. Maybe, three? :-)))"

Then, about three years later, researchers at places like Google and OpenAI began building a new kind of artificial intelligence.

Write me a sonnet

On a recent morning, I asked ChatGPT the same questions that Turing had laid out in his 1950 paper. It instantly generated a poem about the Forth Bridge:

Its red paint gleams in the morning sun

A sight to behold, for all to see

Its majesty and grandeur never done

Then it correctly added 34,957 and 70,764. It did not need 30 seconds to do so. When I laid out the end of a chess game as Turing did, it responded with typically clear, concise, confident prose. It seemed to understand the situation.

But it did not. It mistook the end of the game for the beginning. “I would move my rook to R2,” it said. “It is generally a good idea to try to develop your pieces (move them out from their starting positions) as quickly as possible in chess.”

ChatGPT is what researchers call a neural network, a mathematical system loosely modeled on the network of neurons in the brain. This is the same technology that translates between English and Spanish on services like Google Translate and identifies pedestrians as self-driving cars weave through city streets.
A neural network learns skills by analyzing data. By pinpointing patterns in thousands of photos of stop signs, for example, it can learn to recognize a stop sign.

Five years ago, Google, OpenAI and other A.I. labs started designing neural networks that analyzed enormous amounts of digital text, including books, news stories, Wikipedia articles and online chat logs. Researchers call them “large language models.” Pinpointing billions of distinct patterns in the way people connect words, letters and symbols, these systems learned to generate their own text.
They can create tweets, blog posts, poems, even computer programs. They can carry on a conversation — at least up to a point. And as they do, they can seamlessly combine far-flung concepts. You can ask them to rewrite Queen’s pop operetta, “Bohemian Rhapsody,” so that it rhapsodizes about the life of a postdoc academic researcher, and they will.

“They can extrapolate,” said Oriol Vinyals, senior director of deep learning research at the London lab DeepMind, who has built groundbreaking systems that can juggle everything from language to three-dimensional video games. "They can combine concepts in ways you would never anticipate."

Researchers, businesses and other early adopters have been testing these systems for years. Initially, they were difficult to use. And they spat out as much nonsense as coherent language. But with ChatGPT, OpenAI has refined the technology.
As people tested an early version of the system, OpenAI asked them to rate its responses, specifying whether they were convincing or truthful or useful. Then, through a technique called reinforcement learning, the lab used these ratings to hone the system and more carefully define what it would and would not do.
The result is a chatbot geared toward answering individual questions — the very thing that Turing envisioned. Google, Meta and other organizations have built bots that operate in similar ways.

The trouble is that while their language skills are shockingly impressive, the words and ideas are not always backed by what most people would call reason or common sense. The systems write recipes with no regard for how the food will taste. They make little distinction between fact and fiction. They suggest chess moves with complete confidence even when they do not understand the state of the game.
Because they are trained on data from across the internet, there are an infinite number of situations where they seem to get things right while actually getting them very wrong.

Dr. Sutskever of OpenAI compares these bots to the automated driving service that Tesla calls Full Self Driving. This experimental technology can drive itself on city streets. But you — the human driver — are required to keep your eyes on the road and take control of the car at any moment.

"It does everything. It turns and it stops and it sees all the pedestrians," he said. "And yet you have to intervene fairly frequently."

ChatGPT does question-and-answer, but it tends to break down when you take it in other directions. Franz Broseph can negotiate Diplomacy moves for a few minutes, but if each round of negotiations had been a little longer, Mr. De Graaff might well have realized it was a bot. And if Franz Broseph were dropped into any other situation — like answering tech support calls — it would be useless.

A New Test
Six months before releasing its chatbot, OpenAI unveiled a tool called DALL-E.

A nod to both "WALL-E," the 2008 animated movie about an autonomous robot, and Salvador Dalí, the Surrealist painter, this experimental technology lets you create digital images simply by describing what you want to see. This is also a neural network, built much like Franz Broseph or ChatGPT. The difference is that it learned from both images and text. Analyzing millions of digital images and the captions that described them, it learned to recognize the links between pictures and words.

This is what's known as a multimodal system. Google, OpenAI and other organizations are already using similar methods to build systems that can generate video of people and objects. Start-ups are building bots that can navigate software apps and websites on a user's behalf.

These are not systems that anyone can properly evaluate with the Turing test — or any other simple method. Their end goal is not conversation.

Researchers at Google and DeepMind, which is owned by Google's parent company, are developing tests meant to evaluate chatbots and systems like DALL-E, to judge what they do well, where they lack reason and common sense, and more. One test shows videos to artificial intelligence systems and asks them to explain what has happened. After watching someone tinker with an electric shaver, for instance, the A.I. must explain why the shaver did not turn on.

These tests feel like academic exercises — much like the Turing test. We need something that is more practical, that can really tell us what these systems do well and what they cannot, how they will replace human labor in the near term and how they will not.

We could also use a change in attitude. "We need a paradigm shift — where we no longer judge intelligence by comparing machines to human behavior," said Oren Etzioni, professor emeritus at the University of Washington and founding chief executive of the Allen Institute for AI, a prominent lab in Seattle.

Turing's test judged whether a machine could imitate a human. This is how artificial intelligence is typically portrayed — as the rise of machines that think like people. But the technologies under development today are very different from you and me. They cannot deal with concepts they have never seen before. And they cannot take ideas and explore them in the physical world.

ChatGPT made that clear. As more users experimented with it, they showed off its abilities and limitations. One Twitter user asked ChatGPT what letter came next in the sequence O T T F F S S, and it gave the correct answer (E). But it also told him the wrong reason it was correct, failing to realize that these are the first letters in the numbers 1 to 8.

At the same time, there are many ways these bots are superior to you and me. They do not get tired. They do not let emotion cloud what they are trying to do. They can instantly draw on far larger amounts of information. And they can generate text, images and other media at speeds and volumes we humans never could.

Their skills will also improve considerably in the coming years.

Researchers can rapidly hone these systems by feeding them more and more data. The most advanced systems, like ChatGPT, require months of training, but over those months, they can develop skills they did not exhibit in the past.
"We have found a set of techniques that scale effortlessly," said Raia Hadsell, senior director of research and robotics at DeepMind. "We have a simple, powerful approach that continues to get better and better."

The exponential improvement we have seen in these chatbots over the past few years will not last forever. The gains may soon level out. But even then, multimodal systems will continue to improve — and master increasingly complex skills involving images, sounds and computer code. And computer scientists will combine these bots with systems that can do things they cannot. ChatGPT failed Turing's chess test. But we knew in 1997 that a computer could beat the best humans at chess. Plug ChatGPT into a chess program, and the hole is filled.

In the months and years to come, these bots will help you find information on the internet. They will explain concepts in ways you can understand. If you like, they will even write your tweets, blog posts and term papers.

They will tabulate your monthly expenses in your spreadsheets. They will visit real estate websites and find houses in your price range. They will produce online avatars that look and sound like humans. They will make mini-movies, complete with music and dialogue.

"This will be the next step up from Pixar — superpersonalized movies that anyone can create really quickly," said Bryan McCann, former lead research scientist at

Salesforce, who is exploring chatbots and other A.I. technologies at a start-up called You.com.

As ChatGPT and DALL-E have shown, this kind of thing will be shocking, fascinating and fun. It will also leave us wondering how it will change our lives. What happens to people who have spent their careers making movies? Will this technology flood the internet with images that seem real but are not? Will their mistakes lead us astray?

“All the President’s Men,” Carl Bernstein and Bob Woodward’s classic tale of uncovering Watergate, tells a story about a history paper that Mr. Woodward wrote as a freshman at Yale. After reading countless documents that described King Henry IV standing barefoot in the snow for days as he waited to beg forgiveness from Pope Gregory in 1077, Mr. Woodward included the anecdote in his paper. His professor gave the paper a failing grade, explaining that no human being could stand barefoot in the snow for so long without his feet freezing off.

“The divine right of kings did not extend to overturning the laws of nature and common sense,” the professor said.

Drawing from endless documents about King Henry’s visit to Canossa, ChatGPT might well make the same mistake. You must play the professor.

Certainly, these bots will change the world. But the onus is on you to be wary of what these systems say and do, to edit what they give you, to approach everything you see online with skepticism. Researchers know how to give these systems a wide range of skills, but they do not yet know how to give them reason or common sense or a sense of truth.

That still lies with you.