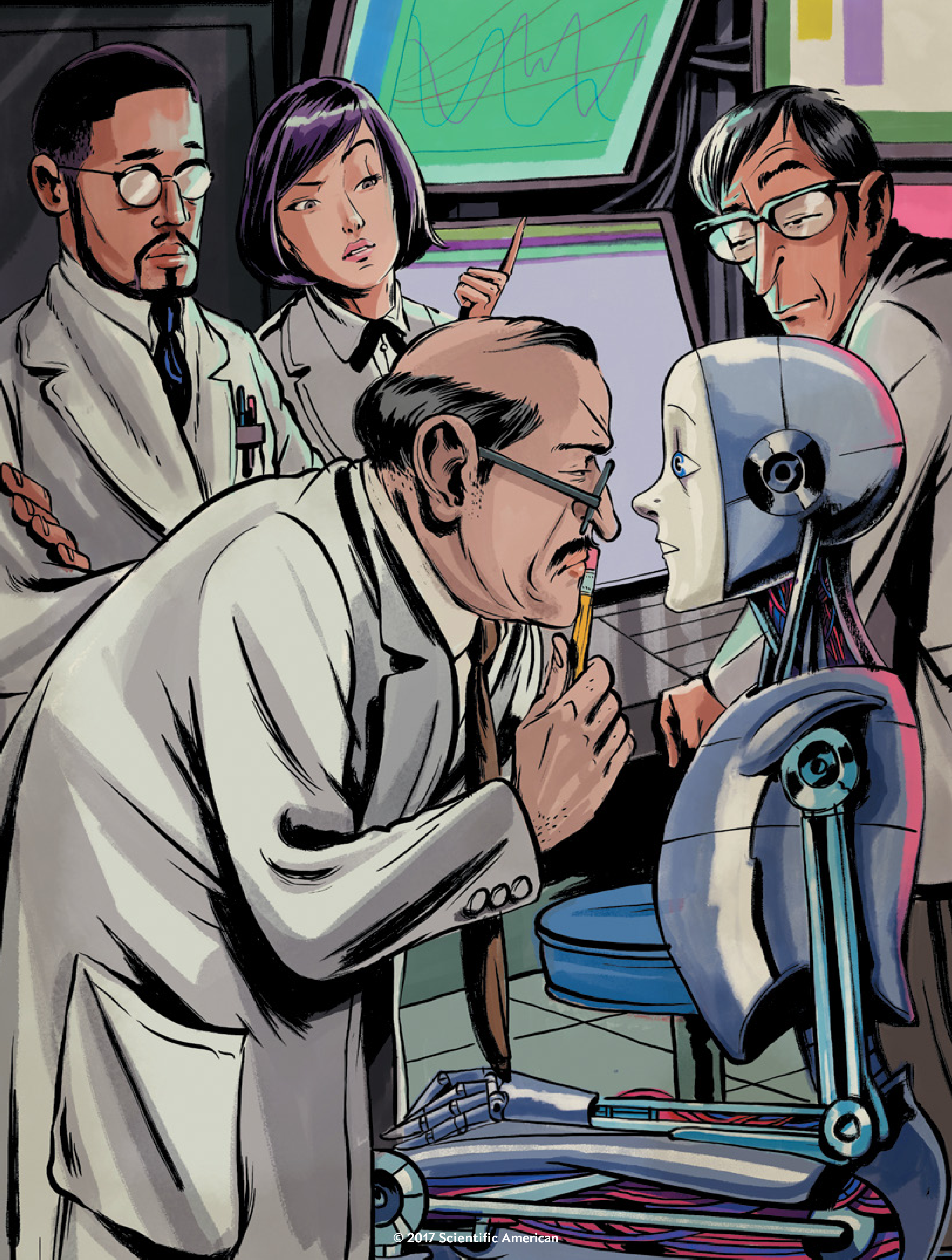


ROBOTICS

AM I HUMAN?

RESEARCHERS NEED NEW WAYS TO DISTINGUISH ARTIFICIAL INTELLIGENCE FROM THE NATURAL KIND

By Gary Marcus

GETTY IMAGES (*fingerprint*)

Illustrations by Zohar Lazar

>>**In 1950** Alan Turing devised a thought experiment that has since been revered as the ultimate test of machine intelligence. He called it the "imitation game," but most people know it as the Turing test. Anticipating what we now call chat bots—computer programs that masquerade as humans—Turing envisioned a contest in which a machine tries to trick an interrogator into believing it is human, answering questions about poetry and deliberately making mistakes about arithmetic. Today, in the eyes of the general public, the Turing test is often seen as a kind of a Rubicon, a measure of whether machines have truly arrived. But it shouldn't be: the Rubicon can be passed, though for the wrong reasons. It is now possible to build machines that fool people, at least for brief periods—but the victories are fleeting, and they do not seem to be carrying us even close to genuine intelligence.

Gary Marcus is director of Uber AI Labs and a professor of psychology and neural science at New York University. His most recent book, co-edited with Jeremy Freeman, is *The Future of the Brain* (Princeton University Press, 2014).

The problem is that Turing's test is too easily gamed. One can "win" simply by being deceptive or feigning ignorance. As Turing himself partly anticipated, the trick to winning is mostly *not* to answer the questions. If a judge asks a machine, "Do you have feelings?" the machine must lie to win. But that odd twist is not an exception—it is the rule. The way to win the Turing test is to deceive.

In 2014, for example, a chat bot named "Eugene Goostman" supposedly "won" the Turing test by pretending to be a sarcastic 13-year-old boy from Odessa, Ukraine, who never answered questions directly. Some judges were fooled (about a third, enough to count as a shallow victory). But so what? The judges got only a few minutes to conduct their interviews, and they undoubtedly would have ferreted out the truth soon enough. In reality, Eugene knew—and understood—almost nothing. Consider this dialogue, which computer scientist Scott Aaronson had with the bot:

SCOTT: *Which is bigger, a shoebox or Mount Everest?*

>> **EUGENE:** *I can't make a choice right now. I should think it out later. And I forgot to ask you where you are from …*

SCOTT: *How many legs does a camel have?*

>> **EUGENE:** *Something between 2 and 4. Maybe, three? :-))) By the way, I still don't know your specialty—or, possibly, I've missed it?*

SCOTT: *How many legs does a millipede have?*

>> **EUGENE:** *Just two, but Chernobyl mutants may have them up to five. I know you are supposed to trick me.*

SCOTT: *No, I need to know that you're not a chatbot. Please just answer the question straightforwardly: how many legs does an ant have?*

>> **EUGENE:** *Something between 2 and 4. Maybe, three? :-))) Oh, what a fruitful conversation ;-).*

Clearly, Turing's original test has not withstood the test of time.

Not long after Goostman hit the media, I sug-

IN BRIEF

In the mind of the public, Alan Turing's "imitation game," in which a machine tries to convince an interrogator that it is human, has long been considered the ultimate test of artificial intelligence.

But Turing's test has not aged well. Passing it is more a matter of deception than of true intelligence. AI experts argue that the time has come to replace Turing's test with a battery of events that will assess machine intelligence from many different perspectives.

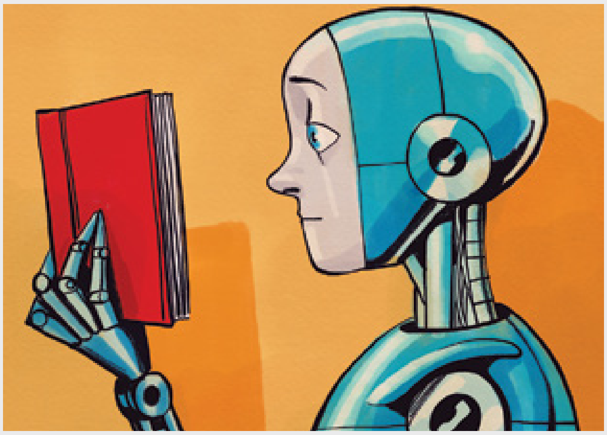
A truly intelligent machine should be able to understand ambiguous statements, build a piece of flat-packed furniture, pass a fourth-grade science test, and more. The difficulty of these tasks underscores the fact that, hype aside, human-level artificial intelligence remains very far in the future.

THE NEW TURING TESTS

AI researchers are developing a variety of tests to replace Alan Turing's 67-year-old "imitation game." Here's a look at four different approaches.

By John Pavlus

TEST 01



Winograd Schema Challenge

Named after pioneering AI researcher Terry Winograd, a "Winograd schema" is a simple but ambiguously worded natural-language question. Answering correctly requires a "commonsense" understanding of how agents, objects and cultural norms influence one another in the real world.

Winograd's first schema, which he wrote in 1971, sets a scene ("The city councilmen refused the demonstrators a permit because they feared violence") and then poses a simple question about it ("Who feared violence?"). This is known as a pronoun disambiguation problem (PDP): in this case, there is ambiguity about whom the word "they" refers to. But Winograd schemas are subtler than most PDPs because the meaning of the sentence can be reversed by changing a single word. (For example: "The city councilmen refused the demonstrators a permit because they *advocated* violence.") Most people use "common sense" or "world knowledge" about typical relationships between city councilmen and demonstrators to resolve the problem. This challenge uses an initial round of PDPs to weed out less intelligent systems; ones that make the cut are given true Winograd schemas.

PROS: Because Winograd schemas rely on knowledge that computers lack reliable access to, the challenge is robustly Google-proof—that is, hard to game with Internet searches.

CONS: The pool of usable schemas is relatively small. "They're not easy to come up with," says Ernest Davis, a professor of computer science at New York University.

DIFFICULTY LEVEL: High. In 2016 four systems competed to answer a set of 60 Winograd schemas. The winner got only 58 percent of the questions correct—far short of the 90 percent threshold that researchers consider a passing grade.

WHAT IT IS USEFUL FOR: Distinguishing comprehension from mere simulations of it. "[Apple's digital assistant] Siri has no understanding of pronouns and cannot disambiguate," explains Leora Morgenstern, a researcher at Leidos who worked on the Winograd Schema Challenge with Davis. That means "you really can't carry on a dialogue [with the system], because you're always referring to something previous in the conversation."

TEST 02

Standardized Testing for Machines

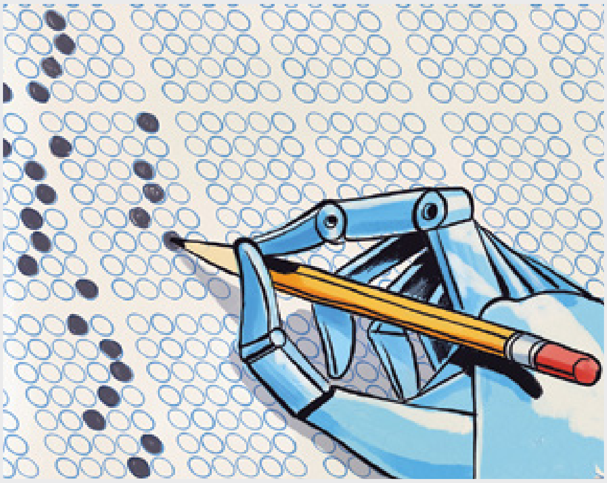
AI would be given the same standardized, written educational tests that we give to elementary and middle school students, without any handholding. The method would assess a machine's ability to link facts together in novel ways through semantic understanding. Much like Turing's original imitation game, the scheme is ingeniously direct. Simply take any sufficiently rigorous standardized test (such as the multiple-choice parts of New York State's fourth-grade Regents science exams), equip the machine with a way of ingesting the test material (such as natural-language processing and computer vision) and let 'er rip.

PROS: Versatile and pragmatic. Unlike Winograd schemas, standardized test material is cheap and abundant. And because none of the material is adapted or preprocessed for the machine's benefit, test questions require a wealth of versatile, commonsense world knowledge just to parse, much less answer correctly.

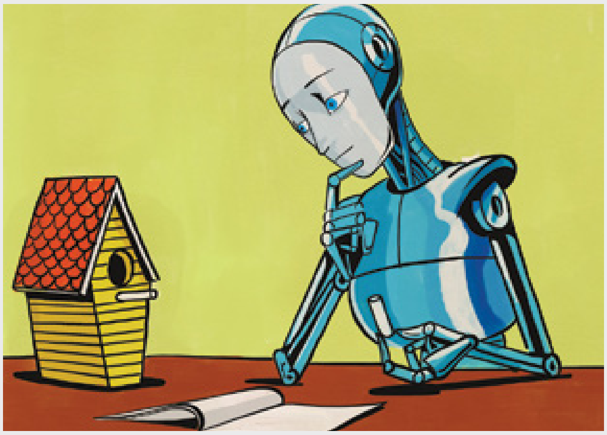
CONS: Not as Google-proof as Winograd schemas, and as with humans, the ability to pass a standardized test does not necessarily imply "real" intelligence.

DIFFICULTY LEVEL: Moderately high. A system called Aristo, designed by the Allen Institute for Artificial Intelligence, achieves an average 75 percent score on the fourth-grade science exams that it has not encountered before. But this is only on multiple-choice questions without diagrams. "No system to date comes even close to passing a full 4th grade science exam," the Allen Institute researchers wrote in a technical paper published in *AI Magazine*.

WHAT IT IS USEFUL FOR: Administering reality checks. "Fundamentally, we can see that no program can get above 60 percent on an eighth-grade science test—but at the same time, we might read in the news that IBM's Watson is going to medical school and solving cancer," says Oren Etzioni, CEO of the Allen Institute for Artificial Intelligence. "Either IBM had some startling breakthrough, or perhaps they're getting a little bit ahead of themselves."



TEST 03



Physically Embodied Turing Test

Most tests for machine intelligence focus on cognition. This test is more like shop class: an AI has to physically manipulate real-world objects in meaningful ways. The test would comprise two tracks. In the construction track, a physically embodied AI—a robot, essentially—would try to build a structure from a pile of parts using verbal, written and illustrated instructions (imagine assembling IKEA furniture). The exploration track would require the robot to devise solutions to a set of open-ended but increasingly creative challenges using toy blocks (such as "build a wall," "build a house," "attach a garage to the house"). Each track would culminate with a communication challenge in which the robot would be required to "explain" its efforts. The test could be given to individual robots, groups of robots or robots collaborating with humans.

PROS: The test integrates aspects of real-world intelligence—specifically, perception and action—that have been historically ignored or underresearched. Plus, the test is essentially impossible to game: "I don't know how you would, unless someone figured out a way to put instructions for how to build anything that's ever been built on the Internet," says Ortiz of Nuance.

CONS: Cumbersome, tedious and difficult to automate without having machines do their construction in virtual reality. Even then, "a roboticist would say that [virtual reality] is still only an approximation," Ortiz says. "In the real world, when you pick up an object, it might slip, or there might be a breeze to deal with. It's hard for a virtual world to faithfully simulate all those nuances."

DIFFICULTY LEVEL: Science-fictional. An embodied AI that can competently manipulate objects *and* coherently explain its actions would essentially behave like a droid from *Star Wars*—well beyond the current state of the art. "To execute these tasks at the level at which children can do them routinely is an enormous challenge," Ortiz says.

WHAT IT IS USEFUL FOR: Imagining a path to integrating the four strands of artificial intelligence—perception, action, cognition and language—that specialized research programs tend to pursue separately.

TEST 04

I-Athlon

In a battery of partially or completely automated tests, an AI is asked to summarize the contents of an audio file, narrate the storyline of a video, translate natural language on the fly and perform other tasks. The goal is to create an objective intelligence score. Automation of testing and scoring—without human supervision—is the hallmark of this scheme. Removing humans from the process of evaluating machine intelligence may seem ironic, but Murray Campbell, an AI researcher at IBM (and a member of the team that developed Deep Blue), says it is necessary to ensure efficiency and reproducibility. Establishing an algorithmically generated intelligence score for AIs would also free researchers from relying on human intelligence—"with all its cognitive biases," Campbell notes—as a yardstick.

PROS: Objectivity, at least in theory. Once I-Athlon judges decided on how to score each test and weight the results, computers would do the actual scoring and weighting. Judging the results should be as cut-and-dried as reviewing an Olympic photo finish. The variety of tests would also help identify what the IBM researchers call "broadly intelligent systems."

CONS: Inscrutability, potentially. I-Athlon algorithms might give high marks to AI systems that operate in ways that researchers do not fully understand. "It is quite possible that some decisions of advanced AI systems will be very difficult to explain [to humans] in a concise and understandable way," Campbell admits. This so-called black box problem is already becoming an issue for researchers working with convolutional neural networks.

DIFFICULTY LEVEL: It depends. Current systems could perform quite well on some potential I-Athlon events, such as image understanding or language translation. Others, such as explaining the contents of a video narrative or drawing a diagram from a verbal description, are still in the realm of sci-fi.

WHAT IT IS USEFUL FOR: Reducing the impact of human cognitive biases on the work of measuring machine intelligence and quantifying—rather than simply identifying—performance.

John Pavlus is a frequent *Scientific American* contributor.

gested an alternative test, designed to push toward real intelligence rather than just dubious evasion. In a *New Yorker* blog post, I proposed that Turing's test be dumped in favor of a more robust comprehension challenge—"a Turing Test for the twenty-first century."

The goal, as I described it then, was to "build a computer program that can watch any arbitrary TV program or YouTube video and answer questions about its content—'Why did Russia invade Crimea?' or 'Why did Walter White consider taking a hit out on Jessie?'" The idea was to eliminate the trickery and focus on whether systems could actually comprehend the materials to which they were exposed. Programming computers to make wisecracks might not bring us closer to true artificial intelligence, but programming them to engage more deeply in the things that they see might.

Francesca Rossi, then president of the International Joint Conferences on Artificial Intelligence, read my proposal and suggested we work together to make this updated Turing test a reality. Together we enlisted Manuela Veloso, a roboticist at Carnegie Mellon University and former president of the Association for the Advancement of Artificial Intelligence, and the three of us began to brainstorm. Initially we focused on finding a single test that could replace Turing's. But we quickly turned to the idea of *multiple* tests because just as there is no single test of athletic prowess, there cannot be one ultimate test of intelligence.

We also decided to get the AI community as a whole involved. In January 2015 we gathered some 50 leading researchers in Austin, Tex., to discuss a refresh of the Turing test. Over a full day of presentations and discussion, we converged on the notion of a competition with multiple events.

One of those events, the Winograd Schema Challenge, named for AI pioneer Terry Winograd (mentor to Google's Larry Page and Sergey Brin), would subject machines to a test in which language comprehension and common sense intersect. Anyone who has ever tried to program a machine to understand language has quickly realized that virtually every sentence is ambiguous, often in multiple ways. Our brain is so good at comprehending language that we do not usually notice. Take the sentence "The large ball crashed right through the table because it was made of Styrofoam." Strictly speaking, the sentence is ambiguous: the word "it" could refer to the table or the ball. Any human listener will realize that "it" must refer to the table. But that requires tying knowledge of materials science with language comprehension—something that remains far out of reach for machines. Three experts, Hector Levesque, Ernest Davis and Leora Morgenstern, have already developed a test around sentences like these, and speech-recognition company Nuance Communications is offering a cash prize of $25,000 to the first system to win.

Our hope is to include many others, too. A Comprehension Challenge in which machines are tested on their ability to understand images, videos, audio and text would be a natural component. Charles Ortiz, Jr., director of the Laboratory for Artificial Intelligence and Natural Language Processing at Nuance, proposed a Construction Challenge that would test perception and physical action—two important elements of intelligent behavior that were entirely absent from the original Turing test. And Peter Clark of the Allen Institute for Artificial Intelligence proposed giving machines the same standardized tests of science and other disciplines that schoolchildren take.

Aside from the tests themselves, conference attendees discussed guidelines for what counts as a good test. Guruduth Banavar and his colleagues at IBM, for example, emphasized that the tests themselves should be computer-generated. Stuart Shieber of Harvard University emphasized transparency: if the events are to push the field forward, awards should be given only to systems that are open—available to the AI community as a whole—and replicable.

When will machines be able to rise to the challenges that we have set? Nobody knows. But people are already taking some of the events seriously, and that could matter for the world. A robot that has mastered the Construction Challenge could, for example, set up temporary camps for displaced people—on Earth or distant planets. A machine that could pass the Winograd Schema Challenge and a fourth-grade biology exam, for example, would bring us closer to the dream of machines that can integrate the vast literature on human medicine, perhaps a vital first step toward curing cancer or deciphering the brain. AI, like every field, needs clear goals. The Turing test was a nice start; now it is time to build a new generation of challenges. SA

Just as there is no single test of athletic prowess, there cannot be one ultimate test of intelligence.

MORE TO EXPLORE

Computing Machinery and Intelligence. A. M. Turing in *Mind*, Vol. 59, No. 235, pages 433-460; October 1950.
What Comes after the Turing Test? Gary Marcus in *New Yorker*. Published online June 9, 2014. www.newyorker.com/tech/elements/what-comes-after-the-turing-test
Beyond the Turing Test. Special issue of *AI Magazine*, Vol. 37, No. 1; Spring 2016.
The Winograd Schema Challenge: **http://commonsensereasoning.org/winograd.html**

FROM OUR ARCHIVES

Could a Machine Think? Paul M. Churchland and Patricia Smith Churchland; January, 1990.

scientificamerican.com/magazine/sa

SCIENTIFIC AMERICAN ONLINE Watch a talk by Marcus at ScientificAmerican.com/mar2017/turing