The New York Times

Artificial Intelligence Hits the Barrier of Meaning

Machine learning algorithms don't yet understand things the way humans do — with sometimes disastrous consequences.

By Melanie Mitchell
Ms. Mitchell is Professor of Computer Science at Portland State University.

Nov. 5, 2018

You've probably heard that we're in the midst of an A.I. revolution. We're told that machine intelligence is progressing at an astounding rate, powered by "deep learning" algorithms that use huge amounts of data to train complicated programs knows as "neural networks."

Today's A.I. programs can recognize faces and transcribe spoken sentences. We have programs that can spot subtle financial fraud, find relevant web pages in response to ambiguous queries, map the best driving route to almost any destination, beat human grandmasters at chess and Go, and translate between hundreds of languages. What's more, we've been promised that self-driving cars, automated cancer diagnoses, housecleaning robots and even automated scientific discovery are on the verge of becoming mainstream.

The Facebook founder, Mark Zuckerberg, recently declared that over the next five to 10 years, the company will push its A.I. to "get better than human level at all of the primary human senses: vision, hearing, language, general cognition." Shane Legg, chief scientist of Google's DeepMind group, predicted that "human-level A.I. will be passed in the mid-2020s."

As someone who has worked in A.I. for decades, I've witnessed the failure of similar predictions of imminent human-level A.I., and I'm certain these latest forecasts will fall short as well. The challenge of creating humanlike intelligence in machines remains greatly underestimated. Today's A.I. systems sorely lack the essence of human intelligence: *understanding* the situations we

experience, being able to grasp their meaning. The mathematician and philosopher Gian-Carlo Rota famously asked, “I wonder whether or when A.I. will ever crash the barrier of meaning.” To me, this is still the most important question.

The lack of humanlike understanding in machines is underscored by recent cracks that have appeared in the foundations of modern A.I. While today’s programs are much more impressive than the systems we had 20 or 30 years ago, a series of research studies have shown that deep-learning systems can be unreliable in decidedly unhumanlike ways.

I’ll give a few examples.

“The bareheaded man needed a hat” is transcribed by my phone’s speech-recognition program as “The bear headed man needed a hat.” Google Translate renders “I put the pig in the pen” into French as “Je mets le cochon dans le stylo” (mistranslating “pen” in the sense of a writing instrument).

Programs that “read” documents and answer questions about them can easily be fooled into giving wrong answers when short, irrelevant snippets of text are appended to the document. Similarly, programs that recognize faces and objects, lauded as a major triumph of deep learning, can fail dramatically when their input is modified even in modest ways by certain types of lighting, image filtering and other alterations that do not affect humans’ recognition abilities in the slightest.

One recent study showed that adding small amounts of “noise” to a face image can seriously harm the performance of state-of-the-art face-recognition programs. Another study, humorously called “The Elephant in the Room,” showed that inserting a small image of an out-of-place object, such as an elephant, in the corner of a living-room image strangely caused deep-learning vision programs to suddenly misclassify other objects in the image.

Furthermore, programs that have learned to play a particular video or board game at a “superhuman” level are completely lost when the game they have learned is slightly modified (the background color on a video-game screen is changed, the virtual “paddle” for hitting “balls” changes position).

These are only a few examples demonstrating that the best A.I. programs can be unreliable when faced with situations that differ, even to a small degree, from what they have been trained on. The errors made by such systems range from harmless and humorous to potentially disastrous: imagine, for example, an airport security system that won’t let you board your flight because your face is confused with that of a criminal, or a self-driving car that, because of unusual lighting conditions, fails to notice that you are about to cross the street.

Even more worrisome are recent demonstrations of the vulnerability of A.I. systems to so-called adversarial examples. In these, a malevolent hacker can make specific changes to images, sound waves or text documents that while imperceptible or irrelevant to humans will cause a program to make potentially catastrophic errors.

The possibility of such attacks has been demonstrated in nearly every application domain of A.I., including computer vision, medical image processing, speech recognition and language processing. Numerous studies have demonstrated the ease with which hackers could, in principle, fool face- and object-recognition systems with specific minuscule changes to images, put inconspicuous stickers on a stop sign to make a self-driving car's vision system mistake it for a yield sign or modify an audio signal so that it sounds like background music to a human but instructs a Siri or Alexa system to perform a silent command.

These potential vulnerabilities illustrate the ways in which current progress in A.I. is stymied by the barrier of meaning. Anyone who works with A.I. systems knows that behind the facade of humanlike visual abilities, linguistic fluency and game-playing prowess, these programs do not — in any humanlike way — *understand* the inputs they process or the outputs they produce. The lack of such understanding renders these programs susceptible to unexpected errors and undetectable attacks.

What would be required to surmount this barrier, to give machines the ability to more deeply understand the situations they face, rather than have them rely on shallow features? To find the answer, we need to look to the study of human cognition.

Our own understanding of the situations we encounter is grounded in broad, intuitive "common-sense knowledge" about how the world works, and about the goals, motivations and likely behavior of other living creatures, particularly other humans. Additionally, our understanding of the world relies on our core abilities to *generalize* what we know, to form abstract concepts, and to make analogies — in short, to flexibly adapt our concepts to new situations. Researchers have been experimenting for decades with methods for imbuing A.I. systems with intuitive common sense and robust humanlike generalization abilities, but there has been little progress in this very difficult endeavor.

A.I. programs that lack common sense and other key aspects of human understanding are increasingly being deployed for real-world applications. While some people are worried about "superintelligent" A.I., the most dangerous aspect of A.I. systems is that we will trust them too much and give them too much autonomy while not being fully aware of their limitations. As the A.I. researcher Pedro Domingos noted in his book "The Master Algorithm," "People worry that computers will get too smart and take over the world, but the real problem is that they're too stupid and they've already taken over the world."

The race to commercialize A.I. has put enormous pressure on researchers to produce systems that work “well enough” on narrow tasks. But ultimately, the goal of developing *trustworthy* A.I. will require a deeper investigation into our own remarkable abilities and new insights into the cognitive mechanisms we ourselves use to reliably and robustly understand the world. Unlocking A.I.’s barrier of meaning is likely to require a step backward for the field, away from ever bigger networks and data collections, and back to the field’s roots as an interdisciplinary science studying the most challenging of scientific problems: the nature of intelligence.

Melanie Mitchell is Professor of Computer Science at Portland State University and External Professor at the Santa Fe Institute. Her book, “Artificial Intelligence: A Guide for Thinking Humans,” will be published in 2019 by Farrar, Straus, and Giroux.

Follow The New York Times Opinion section on Facebook and Twitter (@NYTopinion).