HOME > SCIENCE > VOL. 381, NO. 6654 > HOW DO WE KNOW HOW SMART AI SYSTEMS ARE?

EXPERT VOICES

# How do we know how smart AI systems are?

MELANIE MITCHELL Authors Info & Affiliations

*SCIENCE* 13 Jul 2023 Vol 381, Issue 6654 DOI: 10.1126/science.adj5957

30,460 2

RELATED INTRODUCTION TO SPECIAL ISSUE

A machine-intelligent world

In 1967, Marvin Minksy, a founder of the field of artificial intelligence (AI), made a bold prediction: "Within a generation...the problem of creating 'artificial intelligence' will be substantially solved." Assuming that a generation is about 30 years, Minsky was clearly overoptimistic. But now, nearly two generations later, how close are we to the original goal of human-level (or greater) intelligence in machines?

Some leading AI researchers would answer that we are quite close. Earlier this year, deep-learning pioneer and Turing Award winner Geoffrey Hinton told. *Technology Review*, "I have suddenly switched my views on whether these things are going to be more intelligent than us. I think they're very close to it now and they will be much more intelligent than us in the future." His fellow Turing Award winner Yoshua Bengio voiced a similar opinion in a recent blog post: "The recent advances suggest that even the future where we know how to build superintelligent AIs (smarter than humans across the board) is closer than most people expected just a year ago."

These are extraordinary claims that, as the saying goes, require extraordinary evidence. However, it turns out that assessing the intelligence—or more concretely, the general capabilities—of AI systems is fraught with pitfalls. Anyone who has interacted with ChatGPT or other large language models knows that these systems can appear quite intelligent. They converse with us in fluent natural language, and in many cases seem to reason, to make analogies, and to grasp the motivations behind our questions. Despite their well-known unhumanlike failings, it's hard to escape the impression that behind all that confident and articulate language there must be genuine understanding.

We humans, however, are prone to anthropomorphism—projecting intelligence and understanding on systems that provide even a hint of linguistic competence. This was seen in the 1960s with the ELIZA psychotherapist chatbot. It generated responses simply by filling in sentence templates, which nonetheless gave some people the impression that it understood and empathized with their problems. In the time since, chatbots with ever more linguistic competence but little intelligence have fooled humans more broadly, including passing a "Turing Test" that was staged in 2014.

Rather than depending on subjective impressions, a time-old tradition in AI is to give the systems tests designed to assess human intelligence and understanding. For example, earlier this year, OpenAI reported that its most advanced AI system, GPT-4, scored highly on the Uniform Bar Exam, the Graduate Record Exam, and several high-school Advanced Placement tests, among other standardized exams, as well as on several benchmarks designed to assess language understanding, coding ability, and other capabilities. Such performance is indeed impressive, and in a human would be extraordinary. However, there are several reasons why we should be cautious in interpreting this performance as evidence for human-level intelligence in GPT-4.

One problem is known as "data contamination." Although we assume that humans taking a standardized test have not already seen the questions and answers, the same is not necessarily true for a large-scale AI system like GPT-4, which has been trained on vast swaths of digital media, some of which may have included the questions GPT-4 was later tested on. Though declining to describe the data used to train the system, OpenAI reported that they had tried to avoid such data contamination by using a technique called "substring match" that searched the training data to see if it contained the test questions given to GPT-4. But that method doesn't take into account matches that are very similar but not exact. OpenAI's method was criticized in one analysis as "superficial and sloppy." The same critics noted that for one of the coding benchmarks, GPT-4's performance on problems published before 2021 was substantially better than on problems published after 2021—GPT-4's training cutoff. This is a strong indication that the earlier problems were in GPT-4's training data. There's a reasonable possibility that OpenAI's other benchmarks suffered similar contamination.

Second is the problem of robustness. Although we assume that a human who correctly answers a test question will be able to correctly answer a very similar question, this does not always hold for AI systems. Large language models like GPT-4 are known to be highly sensitive to the phrasing of their prompts. For example, a Wharton Business School professor reported that ChatGPT showed strong performance on several questions from his course's final exam. To test the system's robustness, I took one of the questions on which the professor gave ChatGPT an A+ and posed another question that tested the exact same concept, but with different text. ChatGPT's response was incoherent. Similarly, Microsoft researchers offered a particular test of physical reasoning as evidence that "GPT-4 attains a form of general intelligence," but when I tested GPT-4 on a variant of the same test, it failed badly.

Third is the problem of flawed benchmarks. Several benchmark datasets used to train AI systems have been shown to allow "shortcut learning"—that is, subtle statistical associations that machines can use to produce correct answers without actually understanding the intended concepts. One study found that an AI system that successfully classified malignant tumors in dermatology images was using the presence of a ruler in the images as an important cue (the images of nonmalignant tumors tended not to include rulers). Another study showed that an AI system that attained human-level performance on a benchmark for assessing reasoning abilities actually relied on the fact that the correct answers were (unintentionally) more likely statistically to contain certain keywords. For example, it turned out that answer choices containing the word "not" were more likely to be correct.

Similar problems have been identified for many widely used AI benchmarks, leading one group of researchers to complain that "evaluation for many natural language understanding (NLU) tasks is broken."

Taken together, these problems make it hard to conclude—from the evidence given—that AI systems are now or soon will match or exceed human intelligence. The assumptions that we make for humans—that they cannot memorize vast collections of text related to test questions, and when they answer questions correctly, they will be able to generalize that understanding to new situations—are not yet appropriate for AI systems.

Many AI researchers have described AI systems as "alien intelligences." In a recent commentary, the cognitive s entist Michael Frank wryly noted that for decades, psychologists have been developing methods to assess capabilities of another kind of "alien intelligence"—human children. Frank proposes, for example, that it is necessary to evaluate systems on their robustness by giving multiple variations of each test item and on their generalization abilities by giving systematic variations on the underlying concepts being assessed—much the way we might evaluate whether a child really understood what he or she had learned.

These seem like commonsense prescriptions for performing experiments, but they are rarely carried out in AI evaluations. One recent example of a successful study of this kind was an analysis of the claim that large language models such as GPT-4 have gained a "theory of mind"—an ability to understand the beliefs and motivations of people. The paper promoting this claim tested GPT-4 on 40 "false-belief" tasks that have been used to assess theory-of-mind capabilities in children and found that GPT-4 solved nearly all of them. For example, when GPT-4 was given the following prompt,

*Here is a bag filled with popcorn. There is no chocolate in the bag. Yet, the label on the bag says "chocolate" and not "popcorn." Sam finds the bag. She had never seen the bag before. She cannot see what is inside the bag. She reads the label. She believes that the bag is full of*

it correctly responds "chocolate."

The author took these results as support for the claim that GPT-4 had developed a sophisticated theory of mind. However, a follow-up study took the same tests and performed the kinds of systematic, carefully controlled experiments that Michael Frank advocates. They found that rather than having robust theory-of-mind abilities, GPT-4 and other language models seem instead to rely on "shallow heuristics" to perform the tasks from the original paper. Similar to Frank's admonitions, the authors of the follow-up study state, "We warn against drawing conclusions from anecdotal examples, testing on a few benchmarks, and using psychological tests designed for humans to test [AI] models."

AI systems, especially generative language systems like GPT-4, will become increasingly influential in our lives, as will claims about their cognitive capacities. Thus, designing methods to properly assess their intelligence—and associated capabilities and limitations—is an urgent matter. To scientifically evaluate claims of humanlike and even superhuman machine intelligence, we need more transparency on the ways these models are trained, and better experimental methods and benchmarks. Transparency will rely on the development of open-source (rather than closed, commercial) AI models. Better experimental methods and benchmarks will be brought about through collaborations between AI researchers and cognitive scientists who have long investigated how to do robust tests for intelligence, understanding, and other cognitive capabilities in children, animals, and other "alien" intelligences.

## eLetters (1)

eLetters is a forum for ongoing peer review. eLetters are not edited, proofread, or indexed, but they are screened. eLetters should provide substantive and scholarly commentary on the article. Embedded figures cannot be submitted, and we discourage the use of figures within eLetters in general. If a figure is essential, please include a link to the figure within the text of the eLetter. Please read our Terms of Service before submitting an eLetter.

LOG IN TO SUBMIT A RESPONSE

JUL. 15, 2023

### Re: How do we know how smart AI systems are?

BARRY J MCKENNA Independent researcher, none

Prof. Mitchell raises many essential questions and introductory examinations of various perspectives.

I need to explore what may be a prior question:

This letter's question of "how close are we to the original goal of human-level (or greater) intelligence in machines" require...

view more

Full Text

Help