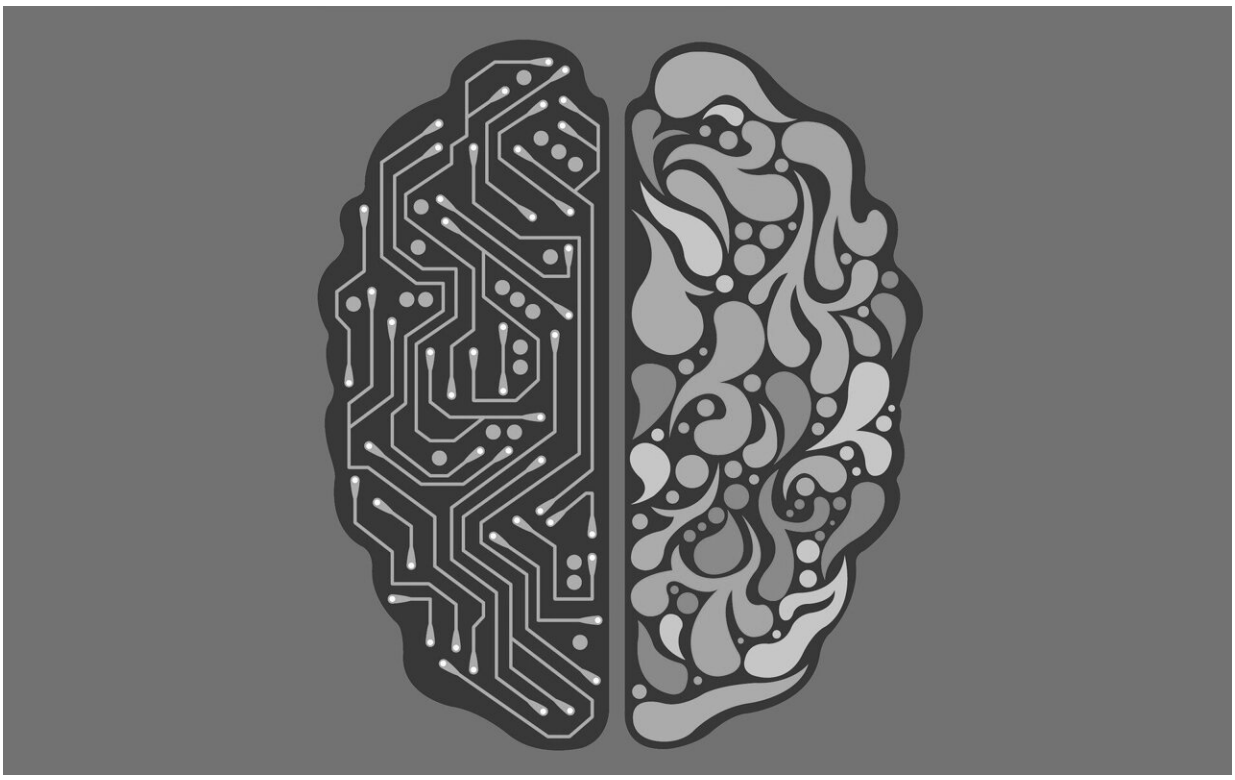
As AI continues to surpass human performance, it's time to reevaluate tests, says expert

April 5 2023, by Shana Lynch



Credit: Pixabay/CC0 Public Domain

How good is AI? According to most of the technical performance benchmarks we have today, it's nearly perfect. But that doesn't mean most artificial intelligence tools work the way we want them to, says

Vanessa Parli, associate director of research programs at the Stanford Institute for Human-Centered AI and a member of the AI Index steering committee.

She cites the current popular example of ChatGPT. "There's been a lot of excitement, and it meets some of these benchmarks quite well," she said. "But when you actually use the tool, it gives incorrect answers, says thing we don't want it to say, and is still difficult to interact with."

In the newest AI Index, published on April 3, a team of independent researchers analyzed over 50 benchmarks in vision, language, speech, and more to find out that AI tools are able to score extremely high on many of these evaluations.

"Most of the benchmarks are hitting a point where we cannot do much better, 80-90% accuracy," she said. "We really need to be thinking about how we, as humans and society, want to interact with AI, and develop new benchmarks from there."

In this conversation, Parli explains more about the benchmarking trends she sees from the AI Index.

What do you mean by benchmark?

A benchmark is essentially a goal for the AI system to hit. It's a way of defining what you want your tool to do, and then working toward that goal. One example is HAI Co-Director Fei-Fei Li's ImageNet, a dataset of over 14 million images. Researchers run their image classification algorithms on ImageNet as a way to test their system. The goal is to correctly identify as many of the images as possible.

What did the AI Index study find regarding these

benchmarks?

We looked across multiple technical benchmarks that have been created over the past dozen years— around vision, around language, etc.—and evaluated the state-of-the-art result in each benchmark year over a year. So, for each benchmark, were researchers able to beat the score from last year? Did they meet it? Or was there no progress at all? We looked at ImageNet, a language benchmark called SUPERGlue, a hardware benchmark called MLPerf, and more; some 50 were analyzed and over 20 made it into the report.

And what did you find in your research?

In earlier years, people were improving significantly on the past year's state of the art or best performance. This year across the majority of the benchmarks, we saw minimal progress to the point we decided not to include some in the report. For example, the best image classification system on ImageNet in 2021 had an accuracy rate of 91%; 2022 saw only a 0.1 percentage point improvement.

So we're seeing a saturation among these benchmarks—there just isn't really any improvement to be made.

Additionally, while some benchmarks are not hitting the 90% accuracy range, they are beating the human baseline. For example, the Visual Question Answering Challenge tests AI systems with open-ended textual questions about images. This year, the top performing model hit 84.3% accuracy. Human baseline is about 80%.

What does that mean for researchers?

The takeaway for me is that perhaps we need newer and more

comprehensive benchmarks to evaluate against. Another way that I think of it is this: Our AI tools right now are not exactly as we would want them to be—they give wrong information, they create sexist imagery. The question becomes, if benchmarks are supposed to help us reach a goal, what is this goal? How do we want to work with AI and how do we want AI to work with us?

Perhaps we need more comprehensive benchmarks—right now, benchmarks mostly test against a single goal. But as we move toward AI tools that incorporate vision, language, and more, do we need benchmarks that help us understand the tradeoffs between accuracy and bias or toxicity, for example? Can we consider more social factors? A lot cannot be measured through quantitative benchmarks. I think this is an opportunity to reevaluate what we want from these tools.

Are researchers already beginning to build better benchmarks?

Being at Stanford HAI, home to the Center for Research on Foundation Models, I can point to HELM. HELM, developed by scholars at CRFM, looks across multiple scenarios and multiple tasks and is more comprehensive than benchmarks we have seen in the past. It considers not only accuracy, but also fairness, toxicity, efficiency, robustness, and more.

That's just one example. But we need more of these approaches. Because benchmarks guide the direction of AI development, they must align more with how we, as humans and as a society, want to interact with these tools.

Provided by Stanford University