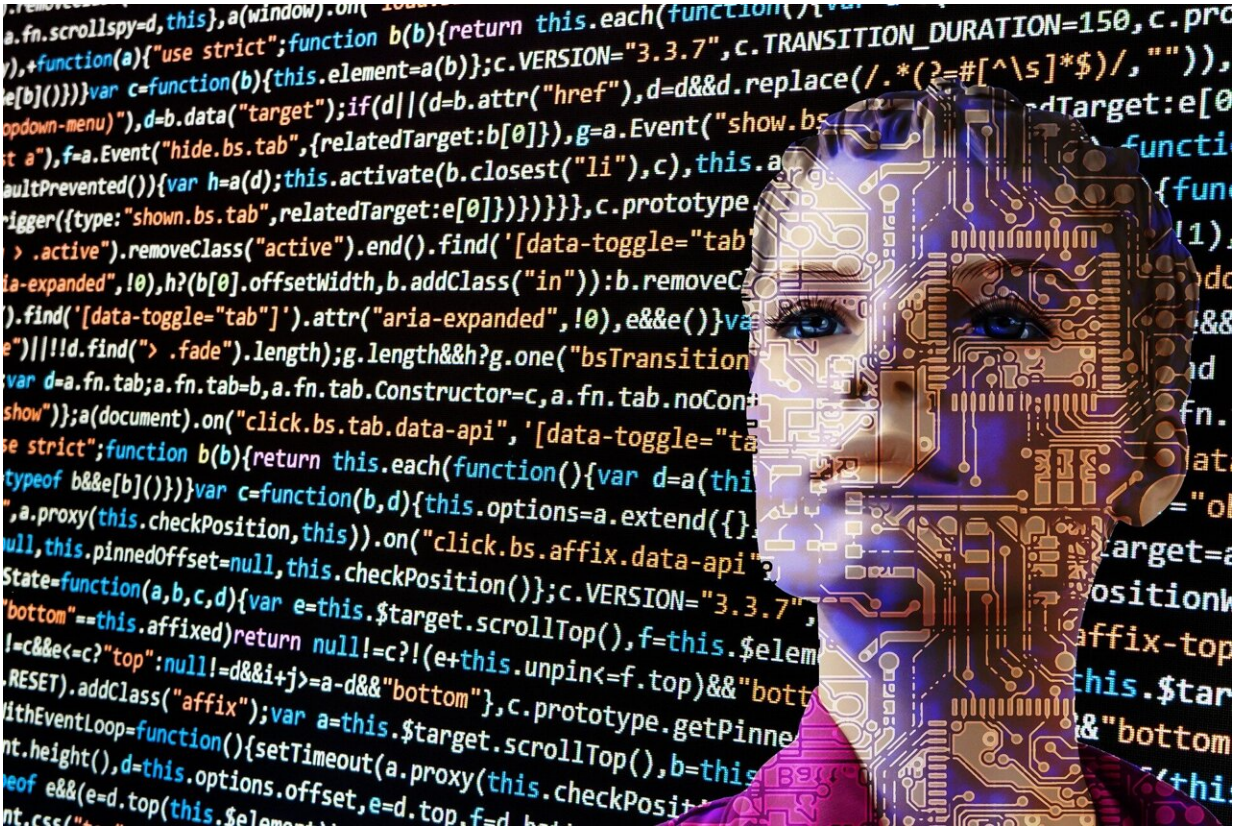
Researchers say AI emergent abilities are just a 'mirage'

May 5 2023, by Peter Grad



Datasets used to train AI algorithms may underrepresent older people. Credit: Pixabay/CC0 Public Domain

There seems to be no end to predictions of storm clouds when computers eventually decide to take matters into their own hands (or should we say,

their own processors).

"The development of artificial intelligence could spell the end of the human race," Stephen Hawking warned.

"[AI] scares the hell out of me. It's capable of vastly more than almost anyone knows, and the rate of improvement is exponential," said OpenAI cofounder Elon Musk.

AI technologies present "profound risks to society and humanity," according to a letter signed earlier this year by more than 1,000 technology leaders urging a moratorium on AI research until more is understood about potential risks.

"We need to be very careful," said Yoshua Bengio, a professor and AI researcher at the University of Montreal.

While not disregarding the promise of tremendous good that AI will bring to a broad range of sectors in industry, economics, education, science, agriculture, medicine and research, media reports are increasingly sounding an alarm over the unintended consequences of this burgeoning disruptive technology.

One area of concern is emergent behavior, defined as a series of unanticipated, unprogrammed interactions within a system stemming from simpler programmed behaviors by individual parts.

Researchers say evidence of such behavior is seen in models that learn languages on their own, when systems trained to play chess and Go generate original strategies to advance, or when robots exhibit variability in motion patterns that were not originally programmed.

"Despite trying to expect surprises, I'm surprised at the things these

models can do," said Google computer scientist Ethan Dyer, responding to an AI experiment in which a computer unexpectedly deduced on its own the title of a movie based on a string of emojis.

But Dyer himself may be surprised to learn that a research team at Stanford University is throwing cold water on reports of emergent behavior.

Ryan Schaeffer, Brando Miranda and Sanmi Koyejo said in a paper posted last week that evidence for emergent behaviors is based on statistics that likely were misinterpreted.

"Our message is that previously claimed emergent abilities … might likely be a mirage induced by researcher analyses," they said.

In their paper posted on the *arXiv* preprint server, the researchers explained that the abilities of large language models are measured by determining the percentage of its correct predictions.

Statistical analyses may be represented in numerous ways. The researchers contend that when results are reported in non-linear, or discontinuous, metrics, they appear to show sharp, unpredictable changes that are erroneously interpreted as indicators of emergent behavior.

However, an alternate means of measuring the identical data using linear metrics shows "smooth, continuous" changes that, contrary to the former measure, reveal predictable—non-emergent—behavior.

The Stanford team added that failure to use large enough samples also contributes to faulty conclusions.

"Existing claims of emergent abilities are creations of the researcher's analyses, not fundamental changes in model behavior on specific tasks,"

the team said.

They added that while methodology in past research likely yielded misleading conclusions, "nothing in this paper should be interpreted as claiming that large language models cannot display emergent abilities," suggesting proper methodology may well reveal such capacities.

"The main takeaway," the researchers said, "is for a fixed task and a fixed model family, the researcher can choose a metric to create an emergent ability or choose a metric to ablate an emergent ability."

Or as one notable commenter stated, "The output of the algorithm is only as good as the parameters which its creators set, meaning there is room for potential bias within the AI itself."

And who was that notable commentator? Microsoft Bing's ChatGPT.

More information: Rylan Schaeffer et al, Are Emergent Abilities of Large Language Models a Mirage?, *arXiv* (2023). DOI: 10.48550/arxiv.2304.15004