What is AI superintelligence? Could it destroy humanity? And is it really almost here?

The Conversation - Australia

October 28, 2024 Monday 7:08 PM EST

Length: 1260 words

Byline: Flora Salim, Professor, School of Computer Science and Engineering, inaugural ***Cisco*** Chair of Digital Transport & AI, UNSW Sydney

Highlight: AI systems more intelligent than humans in some ways already exist - but general-purpose superhuman intelligence is probably still a long way off.

Body

In 2014, the British philosopher Nick Bostrom published a book about the future of artificial intelligence (AI) with the ominous title *Superintelligence: Paths, Dangers, Strategies*. It proved highly influential in promoting the idea that advanced AI systems - "superintelligences" more capable than humans - might one day take over the world and destroy humanity.

A decade later, OpenAI boss Sam Altman says superintelligence may only be "*a few thousand days*" away. A year ago, Altman's OpenAI cofounder Ilya Sutskever set up a team within the company to focus on "*safe superintelligence*", but he and his team have now raised a billion dollars to create *a startup of their own* to pursue this goal.

What exactly are they talking about? Broadly speaking, superintelligence is *anything more intelligent than humans*. But unpacking what that might mean in practice can get a bit tricky.

Different kinds of AI

In my view the most useful way to think about *different levels and kinds of intelligence in AI* was developed by US computer scientist Meredith Ringel Morris and her colleagues at Google.

Their framework lists six levels of AI performance: no AI, emerging, competent, expert, virtuoso and superhuman. It also makes an important distinction between narrow systems, which can carry out a small range of tasks, and more general systems.

A narrow, no-AI system is something like a calculator. It carries out various mathematical tasks according to a set of explicitly programmed rules.

There are already plenty of very successful narrow AI systems. Morris gives the *Deep Blue chess program* that famously defeated world champion Garry Kasparov way back in 1997 as an example of a virtuoso-level narrow AI system.

Some narrow systems even have superhuman capabilities. One example is *Alphafold*, which uses machine learning to predict the structure of protein molecules, and whose creators *won the Nobel Prize in Chemistry* this year.

What about general systems? This is software that can tackle a much wider range of tasks, including things like learning new skills.

A general no-AI system might be something like *Amazon's Mechanical Turk*: it can do a wide range of things, but it does them by asking real people.

Overall, general AI systems are far less advanced than their narrow cousins. According to Morris, the state-of-the-art language models behind chatbots such as ChatGPT are general AI - but they are so far at the "emerging" level (meaning they are "equal to or somewhat better than an unskilled human"), and yet to reach "competent" (as good as 50% of skilled adults).

So by this reckoning, we are still some distance from general superintelligence.

How intelligent is AI right now?

As Morris points out, precisely determining where any given system sits would depend on having reliable tests or benchmarks.

Depending on our benchmarks, an image-generating system such as *DALL-E* might be at virtuoso level (because it can produce images 99% of humans could not draw or paint), or it might be emerging (because it produces errors no human would, such as mutant hands and impossible objects).

There is significant debate even about the capabilities of current systems. One notable 2023 paper *argued* GPT-4 showed "sparks of artificial general intelligence".

OpenAI says its latest language model, *o1*, can "perform complex reasoning" and "rivals the performance of human experts" on many benchmarks.

However, *a recent paper from Apple researchers* found o1 and many other language models have significant trouble solving genuine mathematical reasoning problems. Their experiments show the outputs of these models seem to resemble sophisticated pattern-matching rather than true advanced reasoning. This indicates superintelligence is not as imminent as many have suggested.

Will AI keep getting smarter?

Some people think the rapid pace of AI progress over the past few years will continue or even accelerate. Tech companies are investing *hundreds of billions of dollars* in AI hardware and capabilities, so this doesn't seem impossible.

If this happens, we may indeed see general superintelligence within the "few thousand days" proposed by Sam Altman (that's a decade or so in less scifi terms). Sutskever and his team mentioned a similar timeframe in their *superalignment article*.

Many recent successes in AI have come from the application of a technique called "deep learning", which, in simplistic terms, finds associative patterns in gigantic collections of data. Indeed, *this year's Nobel Prize in Physics* has been awarded to John Hopfield and also the "*Godfather of AI*" Geoffrey Hinton, for their invention of Hopfield Networks and Boltzmann machine, which are the foundation for many powerful deep learning models used today.

General systems such as ChatGPT have relied on data generated by humans, much of it in the form of text from books and websites. Improvements in their capabilities have largely come from increasing the scale of the systems and the amount of data on which they are trained.

However, there *may not be enough human-generated data* to take this process much further (although efforts to use data more efficiently, generate synthetic data, and improve transfer of skills between different domains may bring improvements). Even if there were enough data, some researchers say language models such as ChatGPT are *fundamentally incapable* of reaching what Morris would call general competence.

One recent paper has suggested an essential feature of superintelligence would be *open-endedness*, at least from a human perspective. It would need to be able to continuously generate outputs that a human observer would regard as novel and be able to learn from.

Existing foundation models are not trained in an open-ended way, and existing open-ended systems are quite narrow. This paper also highlights how either novelty or learnability alone is not enough. A new type of open-ended foundation model is needed to achieve superintelligence.

What are the risks?

So what does all this mean for the risks of AI? In the short term, at least, we don't need to worry about superintelligent AI taking over the world.

But that's not to say AI doesn't present risks. Again, Morris and co have thought this through: as AI systems gain great capability, they may also gain greater autonomy. Different levels of capability and autonomy present different risks.

For example, when AI systems have little autonomy and people use them as a kind of consultant - when we ask ChatGPT to summarise documents, say, or let the YouTube algorithm shape our viewing habits - we might face a risk of over-trusting or *over-relying on them*.

In the meantime, Morris points out other risks to watch out for as AI systems become more capable, ranging from people forming parasocial relationships with AI systems to mass job displacement and society-wide ennui.

What's next?

Let's suppose we do one day have superintelligent, fully autonomous AI agents. Will we then face the risk they could concentrate power or act against human interests?

Not necessarily. Autonomy and control *can go hand in hand*. A system can be highly automated, yet provide a high level of human control.

Like many in the AI research community, I believe **safe superintelligence** is feasible. However, building it will be a complex and multidisciplinary task, and researchers will have to tread unbeaten paths to get there.

Flora Salim receives funding from Australian Research Council and *Cisco*. She acknowledges the support from the ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S) (CE200100005).

Load-Date: October 28, 2024