Smarter machines risk creating dumber humans

TECHNOLOGY

John Thornhill

When one of Google's senior researchers asked the company's LaMDA chatbot whether it was a "philosophical zombie" (exhibiting human-like behaviour without having any inner life, consciousness or sentience) it replied: "Of course not." Unconvinced, Blaise Aguera y Arcas asked the AI-enabled chatbot how he could know this was true. "You'll just have to take my word for it. You can't 'prove' you're not a philosophical zombie either," LaMDA answered.

Our machines are becoming smarter — and sassier — at astonishing and unnerving speed. LaMDA is one of a new generation of large language, or foundation, models, which use machine learning techniques to identify patterns of words in vast data sets and automatically replicate them on demand. They operate like speedy auto-complete functions, but with no instinctive or acquired preferences, no memory and no sense of history or identity. "LaMDA is indeed, to use a blunt (if admittedly, humanising) term, bullshitting," Aguera y Arcas wrote.

When OpenAI, a San Francisco-based research company, launched one of the first foundation models, called GPT-3, in 2020 it stunned many users with its ability to generate reams of plausible text at remarkable speed. Since then, such models have become bigger and more powerful, expanding from text to computer code, images and video, too. They are also emerging from sheltered research environments into the wilds of the real world and are increasingly being deployed in marketing, finance, scientific research and healthcare. The critical question is how closely these technological tools should be controlled. The risk is that smarter machines may only make dumber humans.

The technology's positive commercial uses are highlighted by Kunle Olukotun, a Stanford University professor and co-founder of SambaNova Systems, a Silicon Valley start-up that helps clients deploy AI. "The pace of innovation and the size of the models is increasing dramatically," he says. "Just when you thought that we were reaching our limits, people come up with new tricks."

Not only can these new models generate text and images but interpret them too. This enables the same system to learn in different contexts and handle multiple tasks. For example, Hungary's OTP bank is working with the government and SambaNova to deploy AI-powered services across its business. The bank aims to use the technology to add automated agents at its call centres, personalise services to its 17mn retail customers and streamline its internal processes by analysing documents. "Nobody really knows what banking will look like in 10 years' time — or what the technology will look like. But I am 100 per cent sure that AI will play a key role," says Peter Csanyi, OTP's chief digital officer.

> If trained on biased data sets, AI foundation models can devalue the currency of truth and threaten privacy

Some companies that have developed powerful foundation models, such as Google, Microsoft and OpenAI, restrict access to the technology to known users. But others, including Meta and EleutherAI, share it with a broader customer base. There is a tension between allowing outside experts to help detect flaws and bias and preventing more sinister use by the unscrupulous.

Foundation models may be "really exciting and impressive" but are open to abuse because they are "designed to be devious", says Carissa Véliz of Oxford university's Institute for Ethics in AI. If trained on historically biased data sets, foundation models can produce harmful outputs. They can threaten privacy by extracting digital detail about an individual and using bots to reshape online personas. They can also devalue the currency of truth by flooding the internet with fake information.

Véliz makes an analogy with financial systems: "We can trust money so long as there is not too much counterfeit. But if there is more fake money than real money, the system breaks down. We are creating tools and systems we cannot control." That argues for the implementation of randomised control trials for foundation models before release, she says, just as for pharmaceutical drugs.

The Stanford Institute for Human-Centred AI has pushed for the creation of an expert review board to set community norms, share best practice and agree standardised access rules before foundation models are released. Democracy is not just about transparency and openness. It is also about institutional design for collective governance. We are, as the Stanford institute's Rob Reich puts it, in a race between "disruption and democracy".

Until effective collective governance is put in place to control the use of foundation models, it is far from clear that democracy will win.

john.thornhill@ft.com