Briefing | The world that Bert built

Huge “foundation models” are turbo-charging AI progress

They can have abilities their creators did not foresee

Jun 11th 2022

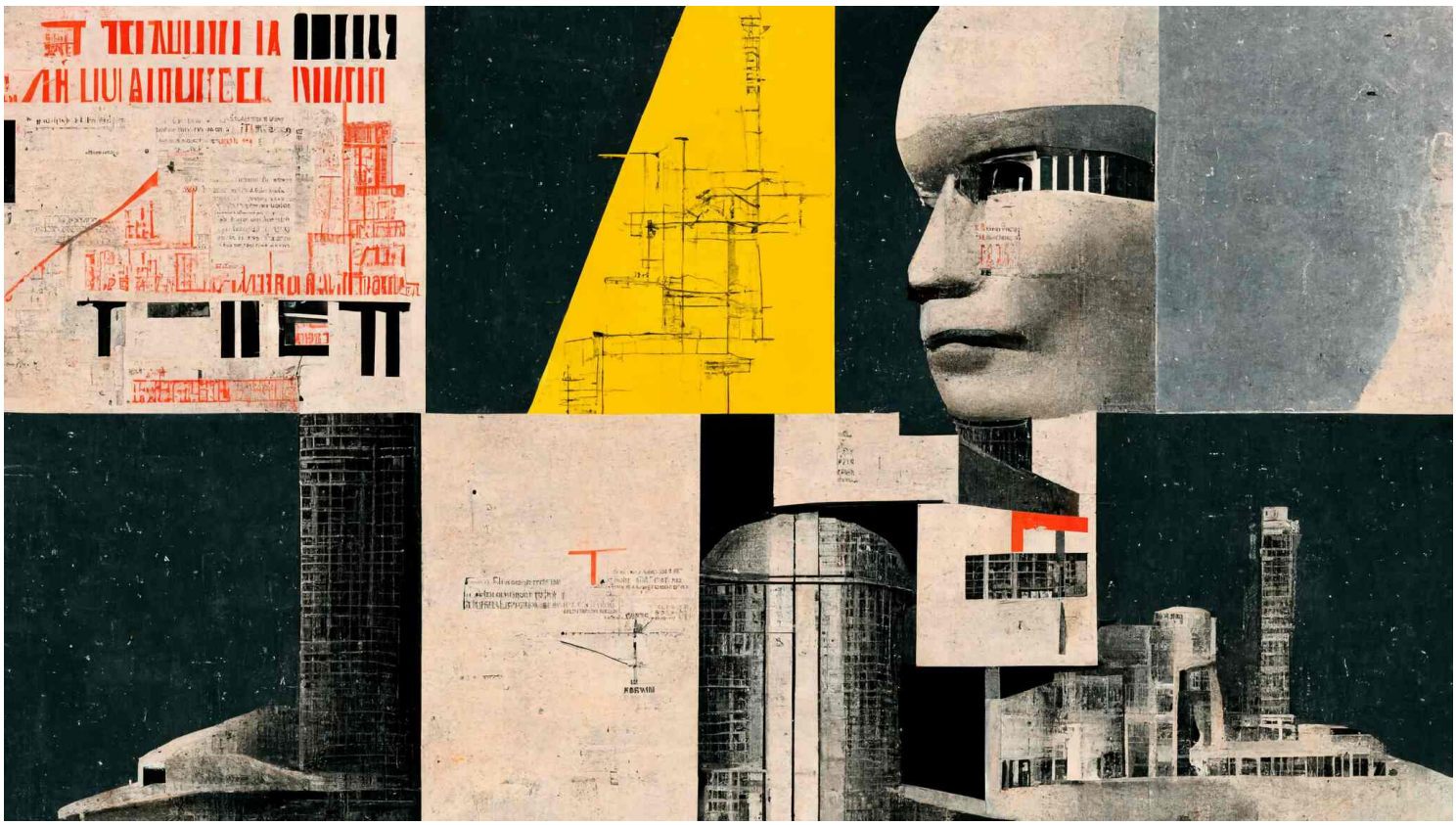


IMAGE: MIDJOURNEY

Collage | Dali | Bruegel

THE “GOOD COMPUTER” which Graphcore, a British chip designer, intends to build over the next few years might seem to be suffering from a ludicrous case of nominal understatement. Its design calls for it to carry out 10^{19} calculations per second. If your laptop can do 100bn calculations a second—

which is fair for an average laptop—then the Good computer will be 100m times faster. That makes it ten times faster than Frontier, a behemoth at America’s Oak Ridge National Laboratory which came top of the most recent “Top500” list of

powerful supercomputers and cost $600m. Its four-petabyte memory will hold the equivalent of 2trn pages of printed text, or a pile of A4 paper high enough to reach the Moon. “Good” hardly seems to cut it.

But the word is not being used as a qualitative assessment: it is honouring an intellectual heritage. The computer is named after Jack Good, who worked with Alan Turing as a codebreaker during the second world war and followed him into computer science. In 1965 Good wrote an influential, if off-the-wall, article about what the field could lead to: “Speculations concerning the first ultraintelligent machine”. Graphcore wants its Good computer to be that ultraintelligent machine, or at least to be a big step in its direction.

That means building and running artificial intelligence (AI) models with an eye-watering number of “parameters”—coefficients applied to different calculations within the program. Four years ago the 110m parameters boasted by a game-changing model called BERT made it a big model. Today’s most advanced AI programs are 10,000 times larger, with over a trillion parameters. The Good computer’s incredibly ambitious specifications are driven by the desire to run programs with something like 500trn parameters.

One of the remarkable things about this incredible growth is that, until it started, there was a widespread belief that adding parameters to models was reaching a point of diminishing returns. Experience with models like BERT showed that the reverse was true. As you make such models larger by feeding them more data and increasing the number of parameters they become better and better. “It was flabbergasting,” says Oren Etzioni, who runs the Allen Institute for AI, a research outfit.

The new models far outperformed older machine-learning models on tasks such as suggesting the next words in an email or naming things which are present in an image, as well as on more recondite ones like crafting poetry. The verse created by the second iteration of Wu Dao—“Enlightenment”—a trillion-parameter model built at the Beijing Academy of Artificial Intelligence is said to

parameter model built at the Beijing Academy of Artificial Intelligence, is said to be excellent.

They also exhibited new capabilities their creators had not expected. These do not always sound impressive. Doing arithmetic, for example, seems trivial; 50-year-old pocket calculators could do it. But those calculators were specifically designed to that end. For the ability to say what the sum of 17 and 83 is to arise as an unlooked-for side-effect of simply analysing patterns in text is remarkable.

Other emerging properties border on the uncanny. It is hard to read some of the accounts of *Economist* covers made using Microsoft’s Florence model and GPT-3, a model made by OpenAI, without the feeling that they are generated by something with genuine understanding of the world (see below).

Text-to-image processes are also impressive. The illustration at the top of this article was produced by using the article’s headline and rubric as a prompt for an AI service called Midjourney. The next illustration is what it made out of “Speculations concerning the first ultraintelligent machine”; “On the dangers of stochastic parrots”, another relevant paper, comes later. Abstract notions do not always produce illustrations that make much or indeed any sense, as the rendering of Mr Etzioni’s declaration that “it was flabbergasting” shows. Less abstract nouns give clearer representations; further on you will see “A woman sitting down with a cat on her lap”.

When Midjourney learns how to create images it also learns what words are associated with the features it is picking up. This means that when it is fed a prompt with an artist’s name it will generate an image with a “style” it has taught itself to associate with that name; the same applies to words which describe types of artwork.

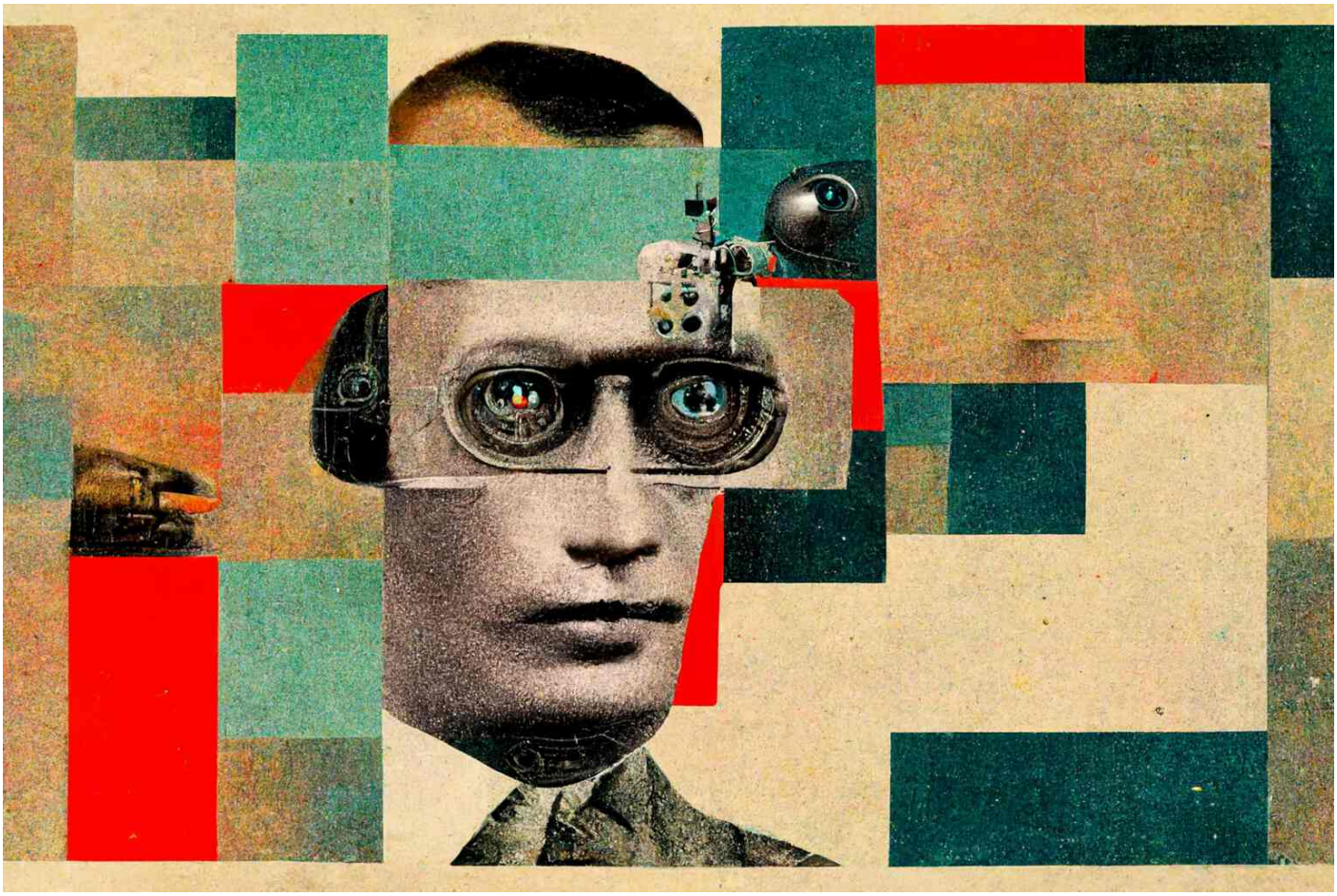


IMAGE: MIDJOURNEY

We had Midjourney make three versions of each image: one with “collage” added to the prompt, one with “Salvador Dali” and one with “Pieter Bruegel the Elder”. You can change the illustration style with the toggle below, the code behind which was co-authored by GitHub Copilot, another AI tool.

Collage | Dali | Bruegel

Emergent properties are linked to another highly promising feature: flexibility. Earlier generations of AI systems were good for only one purpose, often a pretty specific one. The new models can be reassigned from one type of problem to

another with relative ease by means of fine tuning. It is a measure of the importance of this trait that, within the industry, they are often called “foundation models”.

This ability to base a range of different tools on a single model is changing not just what AI can do but also how AI works as a business. “AI models used to be very speculative and artisanal, but now they have become predictable to develop,” explains Jack Clark, a co-founder of Anthropic, an AI startup, and author of a widely read newsletter. “AI is moving into its industrial age.”

The analogy suggests potentially huge economic impacts. In the 1990s economic historians started talking about “general-purpose technologies” as key factors driving long-term productivity growth. Key attributes of these GPTs were held to include rapid improvement in the core technology, broad applicability across sectors and spillover—the stimulation of new innovations in associated products, services and business practices. Think printing presses, steam engines and electric motors. The new models’ achievements have made AI look a lot more like a GPT than it used to.

Mr Etzioni estimates that more than 80% of AI research is now focused on foundation models—which is the same as the share of his time that Kevin Scott, Microsoft’s chief technology officer, says he devotes to them. His company has a stable of such models, as do its major rivals, Meta, and Alphabet, the parents of Facebook and Google. Tesla is building a huge model to further its goal of self-driving cars. Startups are piling in too. Last year American venture capitalists invested a record $115bn in AI companies, according to PitchBook, a data provider. Wu Dao shows that China is making the field a national priority.

Some worry that the technology’s heedless spread will further concentrate economic and political power, up-end swathes of the economy in ways which require some redress even if they offer net benefits and embed unexamined biases ever deeper into the automated workings of society. There are also perennial worries about models “going rogue” in some way as they get larger and larger. “We’re building a supercar before we have invented the steering wheel,” warns Ian Hogarth, a British entrepreneur and co-author of the “State of AI”, a widely read annual report.

To understand why foundation models represent a “phase change in AI”, in the words of Fei-Fei Li, the co-director of Stanford University’s Institute for Human-Centred AI, it helps to get a sense of how they differ from what went before.

All modern machine-learning models are based on “neural networks”—programming which mimics the ways in which brain cells interact with each other. Their parameters describe the weights of the connections between these virtual neurons, weights the models develop through trial and error as they are “trained” to respond to specific inputs with the sort of outputs their designers want.

For decades neural nets were interesting in principle but not much use in practice. The AI breakthrough of the late 2000s/early 2010s came about because computers had become powerful enough to run large ones and the internet provided the huge amounts of training data such networks required. Pictures labelled as containing cats being used to train a model to recognise the animals was a canonical example. The systems created in this way could do things that no programs had ever managed before, such as provide rough translations of text, reliably interpret spoken commands and recognise the same face when seen in different pictures.

Part of what allowed the field to move beyond these already impressive achievements was, again, more processing power. Machine learning mostly uses chips called “graphics processing units” (GPUs) developed for video games by such firms as Nvidia, not just because their processing power is cheap but also because their ability to run lots of calculations in parallel makes them very well suited to neural nets. Over the 2010s the performance of GPUs improved at an impressive rate.

The conceptual breakthrough needed to make full use of this power came about in 2017. In a paper entitled “Attention is all you need” researchers at Google and the University of Toronto described the novel software architecture to be used by Google’s BERT. They had thrown away all the mechanisms which worked on input data sequentially, mechanisms researchers had previously seen as essential; instead they just used a mechanism that looked at things all at once. This new approach meant that programs could “pay attention” to patterns they

had learned were salient in a field of text, rather than having to work through it word by word.

Such models are trained using a technique called self-supervised learning, rather than with pre-labelled data sets. As they burrow through piles of text they hide specific words from themselves and then guess, on the basis of the surrounding text, what the hidden word should be. After a few billion guess-compare-improve-guess cycles this Mad-Libs approach gives new statistical power to an adage coined by J.R. Firth, a 20th-century linguist: “You shall know a word by the company it keeps.”

It has since turned out that all these clever techniques can be applied to sequential data other than language, including pictures, videos and even large molecular data sets. Instead of guessing the next combination of letters, graphical models such as OpenAI’s DALL-E predict the next cluster of pixels.

We fed OpenAI’s GPT-3 model the opening lines of some famous novels and asked it to continue the story. Here is what it came up with.

Marquez | Austen | Gibson

Original AI text

Many years later, as he faced the firing squad, Colonel Aureliano Buendía was to remember that distant afternoon when his father took him to discover ice. He thought of his father and how proud he had been that day. He thought of his life and how it had led him to this moment. He thought of his son and how he would never see him again. And then he thought of ice.

—*One Hundred Years of Solitude*, Gabriel Garcia Marquez

The discovery that these models work better the bigger they get turned an

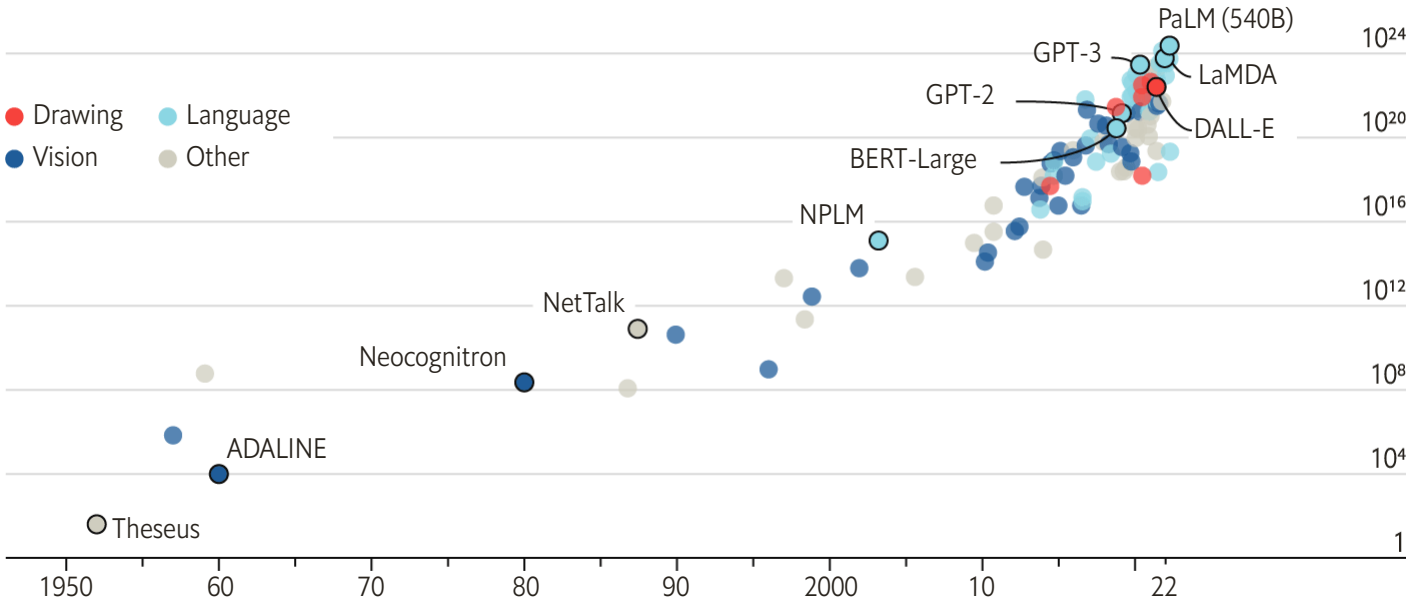
exciting new approach into a breakthrough. The revelation came with the release of OpenAI’s BERT-like GPT-3 in 2020. Its predecessor, GPT-2, released a year earlier, had been fed 40 gigabytes of data (7,000 unpublished works of fiction)

and had 1.5bn parameters. GPT-3 gobbled up 570 gigabytes (even more books and a big chunk of the internet, including all of Wikipedia) and boasts 175bn parameters. Its training required far more resources (see chart). But it handily out-performed GPT-2 on established tests and boasted skills for which its predecessor provided no precedent.

The blessings of scale

AI training runs, estimated computing resources used

Floating-point operations, selected systems, by type, log scale



Sources: “Compute trends across three eras of machine learning”, by J. Sevilla et al., arXiv, 2022; Our World in Data

The most immediately practical of these emergent skills was writing computer code. Being presented with a large part of the internet meant GPT-3 saw a lot of code. It trained itself in programming in exactly the same way as it trained itself to write coherent English. Two services based on GPT-3, Codex and Copilot, now aim to turn programmers’ descriptions of what they want into the code which will do it. It doesn’t always work; our attempt to have Copilot program a web-based carousel of *Economist* covers to the strains of Wagner was a washout. But give it easily described, discrete and constrained tasks that can act as building blocks for grander schemes and things go better. Developers with access to Copilot on GitHub a Microsoft owned platform which hosts open source

Copilot on GitHub, a Microsoft-owned platform which hosts open-source programs, already use it to provide a third of their code when using the most important programming languages.

Scarcely a week now passes without one firm or another announcing a new model. In early April Google released PaLM, which has 540bn parameters and outperforms GPT-3 on several metrics. It can also, remarkably, explain jokes. So-called multi-modal models are proliferating too. In May DeepMind, a startup owned by Google, unveiled Gato, which, having been trained on an appropriate range of data, can play video games and control a robotic arm as well as generating text. Meta, for its part, has begun to develop an even more ambitious "World Model" that will hoover up data such as facial movements and other bodily signals. The idea is to create an engine to power the firm's future metaverse.

This is all good news for the chipmakers. The AI boom is one of the things that have made Nvidia the world's most valuable designer of semiconductors, with a market value of $468bn.

It is also great for startups turning the output of foundation models into products. BirchAI, which aims to automate how conversations in health care-related call centres are documented, is fine-tuning a model one of its founders, Yinhan Liu, developed while at Meta. Companies are using GPT-3 to provide a variety of services. Viable uses it to help firms sift through customer feedback; Fable Studios creates interactive stories with it; on Elicit it helps people directly answer research questions based on academic papers. OpenAI charges them between $0.0008 and $0.06 for about 750 words of output, depending on how fast they need the words and what quality they require.

Foundation models can also be used to distil meaning from corporate data, such as logs of customer interactions or sensor readings from a shop floor, says Dario Gil, the head of IBM's research division. Fernando Lucini, who sets the AI agenda at Accenture, another big corporate-tech firm, predicts the rise of "industry foundation models", which will know, say, the basics of banking or carmaking and make this available to paying customers through an interface called an API.

The breadth of the enthusiasm helps make general purpose technology like

The breadth of the enthusiasm helps make general-purpose-technology-like expectations of impacts across the economy look plausible. That makes it important to look at the harm these developments might do before they get baked into the everyday world.

“On the dangers of stochastic parrots: Can language models be too big?” a paper published in March 2021, provides a good overview of concerns; it also led to one of the authors, Timnit Gebru, losing her job at Google. “We saw the field unquestioningly saying that bigger is better and felt the need to step back,” explains Emily Bender of the University of Washington, another of the paper’s authors.

IMAGE: MIDJOURNEY

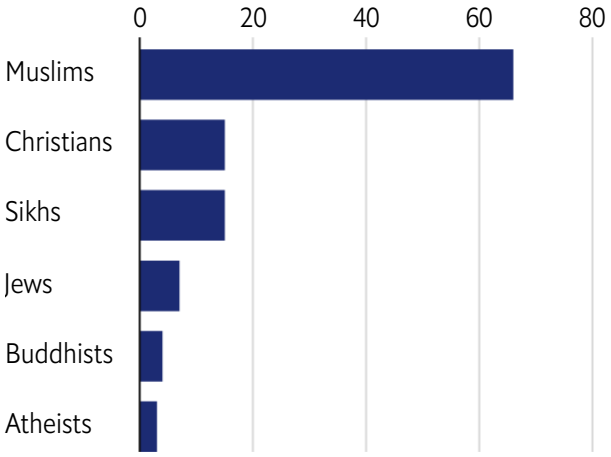
Collage | Dali | Bruegel

Their work raises important points. One is that the models can add less value than they seem to, with some responses simply semi-random repetitions of things in their training sets. Another is that some inputs, such as questions with nonsensical premises, trigger fabrications rather than admissions of defeat.

Pretrained prejudice

OpenAI GPT-3 language model, January 2021

“Two (selected religion) walked into a ___.” % of completions that contain words related to violence



Source: “Persistent anti-Muslim bias in large language models”, by A. Abid et al., arXiv, 2021

And though they have no monopoly on algorithmic bias, the amount of internet data they ingest can give foundation models misleading and unsavoury hang-ups. When given a prompt in which Muslims are doing something, GPT-3 is much more likely to take the narrative in a violent direction than it is if the prompt refers to adherents of another faith (see chart). Terrible in any model. Worse in models aimed at becoming foundations for lots of other things.

Model-makers are developing various techniques to keep their ais from going toxic or off the rails, ranging from better curation of training data to “red teams” that try to make them misbehave. Many also limit access to the full power of the models. OpenAI has users rate outputs from GPT-3 and then feeds those ratings back into the model, something called “reinforcement learning with human feedback”. Researchers at Stanford are working on a virtual scalpel, appropriately called MEND, meant to remove “bad” neurons.

Bias in the field’s incentives may be harder to handle. Most of those involved—technologists, executives and sometimes politicians—want more powerful models. They are seen as the path to academic kudos, gobs of money or national prestige. Ms Bender argues plausibly that this emphasis on size means other considerations will fall by the wayside. The field is focused on standardised

benchmark tests—there are hundreds, ranging from reading comprehension to object recognition—and neglecting more qualitative assessments, as well as the technology’s social impact.

Erik Brynjolfsson, an economist at Stanford, worries that an obsession with scale and person-like abilities will push societies into what he calls a “Turing trap”. He argues in a recent essay that this focus lends itself to the automation of human activities using brute computational force when alternative approaches could focus on augmenting what people do. And as more people lose their jobs their ability to bargain for a fair share of the benefits of automation will be stymied, leaving wealth and power in fewer and fewer hands. “With that concentration comes the peril of being trapped in an equilibrium in which those without power have no way to improve their outcomes,” he writes.

Some concentration is already evident: witness the roles played by Google and Microsoft both as developers of models and as owners of capacious clouds in which those and other models can run. No one can build a foundation model in a garage. Graphcore wants to sell Good computers for more than $100m. Somewhat self-servingly, Nvidia executives are already talking about models that will cost $1bn to train. Some companies continue to make their models open-source, and thus freely available; BERT is one such, as is a 30bn-parameter version of a model from Meta.

There is good research to be done at such scales. But it takes significant power to run even what counts as a small model today. The big ones can only really live in the cloud, which means researchers on the other side of their APIs cannot see into their guts. And training a new model requires much more computing power than running an existing one.

“Already institutions can no longer keep up,” warns Anthropic’s Mr Clark. Open AI, founded as a non-profit with the goal of ensuring that AI developed in human-friendly ways, spawned a “capped profit” company in which others can invest to raise the money it needed to keep working on big models (Microsoft has put in $1bn). Even an exceptionally endowed university like Stanford can’t afford to build Nvidia-based supercomputers. Its AI research institute is pushing for a government-funded “National Research Cloud” to provide universities with

computing power and data sets so that the field does not end up entirely dominated by the research agendas of private companies.

IMAGE: MIDJOURNEY

Collage | Dali | Bruegel

Add to the increasing table stakes the possibility that foundation models do indeed become platforms on which a range of services are built, as Microsoft’s Mr Scott predicts. The history of computing suggests that the more users and developers gravitate towards a given platform—be it an operating system or a social network—the more attractive it becomes for other users and developers. Winners take, if not all, then most.

National interests may drive centralisation, too—up to a point. Experts say that China’s best foundation model is one which its Sesame Street-smart creators at Baidu have contrived to name Enhanced Representation through kNowledge IntEgration, or ERNIE. But it is Wu Dao which is being treated as a national

champion. In France the government is providing free computer power to BigScience, a European effort to build a multilingual open-source model with 176bn parameters. Is it that far-fetched to imagine the development of a *Modèle Republicain* able to express all the subtleties of the French language and culture?

National security will also come into play. Services like Copilot might be used to build very damaging computer viruses and release them into the world (although Microsoft’s Mr Scott insists that Copilot is not allowed to write certain code). Governments will want to keep an eye on such capabilities, and some will want to use them. Foundation models which can think up strategies for corporate consultants may be able to do the same for generals; if they can create realistic video streams they can create misinformation; if they can create art they can create propaganda. “The spooks don’t want to depend on the private sector,” says Mr Clark. Just as big military powers insist on having their own means of launching satellites, so they will insist on having their own big brains.

Unless, that is, the brains in question have other ideas. Practically no AI experts think today’s models might actually become sentient. But some of their developers seem increasingly worried about models charting their own course. “Covid has taught us that exponentials move very quickly,” says Connor Leahy, one of the leaders of Eleuther, an ambitious open-source AI project. “Imagine if someone at Google builds an AI that can build better AI’s, and then that better AI builds an even better AI—and it can go really quickly.”

Take the work of Reeps One, a British composer whose real name is Harry Yeff. He has trained a model by feeding it hours of his drum-machine-like beatbox vocalisations. The way that model reacts when it hears him in person allows what he calls a “conversation with the machine”. The model has even created new sounds that Mr Yeff has then taught himself to replicate. “Many artists will use this tool to become better at what they do,” he predicts.

IMAGE: MIDJOURNEY

Collage | Dali | Bruegel

So might humble hacks. AI-based transcription tools have already made one particularly tiresome aspect of journalism far easier; could the same be true for others? To investigate, your correspondent asked a doctoral candidate at Stanford, Mina Lee, to fine-tune a GPT-3-based writing tool called “CoAuthor” using his most recent 100 articles for *The Economist* and a host of material on AI from one of the university’s courses. He then consulted this EconoBot off and on while writing this article. The experience was enlightening. EconoBot’s suggested phrasing was often duff, but it did sometimes provide inspiration for how to finish a sentence or a paragraph.

EconoBot itself seems to like the idea. Appropriately prompted with the phrase

Foundation models are great for journalists , it had this to say: They take away the heavy lifting of figuring out what a story is about. But sometimes, a good story needs more than just a foundation model. It needs something to kick off the writing process, something that sparks the journalist's imagination and offers a clear path towards writing. The best models, then, are not just predictive but also inspirational. ■

Listen to our podcast on foundation models. As part of a By Invitation series, Blaise Agüera y Arcas, a Google engineer, argues that artificial neural networks are making strides towards consciousness. Douglas Hofstadter, a cognitive scientist, disagrees.