AI's Scale

Donald MacKenzie

An edited version of this draft article was published in the London Review of Books of 5 February 2026: https://www.lrb.co.uk/the-paper/v48/n02/donald-mackenzie/ai-s-scale

The fact-checking links at the end were not published, but are included here as the equivalent of references.

Hyperion is the name that Meta has chosen for a huge AI data centre that the corporation is building in Louisiana. In July, a striking image of Hyperion's footprint superimposed on Manhattan circulated on social media. It stretched across much of the island's width, and from Lower Manhattan to north of Central Park. I assumed that the image had been constructed by a critic of the data centre's scale, but it turned out to have been posted to Threads by Mark Zuckerberg himself.

The imperative to increase scale is embedded deeply in today's AI. Some of its roots lie in an influential interpretation of the field's history. For decades, the neural networks that power many aspects of today's AI – including large language models of the kind that underpin ChatGPT – were marginal to the field, considered less promising than 'symbolic AI' systems, which apply rules roughly akin to those of symbolic logic to systematic bodies of knowledge, often knowledge elicited from human experts. The supporters of neural networks took a different path. They believed that those networks' loose similarity to the brain's interconnected neurons, and their capacity to learn from examples (rather than simply to apply pre-formulated rules) made them a potentially superior route to artificial intelligence. They faced, however, sometimes fierce criticism

from their colleagues, especially MIT's leading AI expert, Marvin Minsky, as well as the fact that early neural networks often simply did not work as well as more mainstream machine-learning techniques.

'Other methods … worked a little bit better', the University of Toronto's Geoffrey Hinton, a leading proponent of neural networks, told *Wired* magazine in 2019. 'That was very depressing … We thought it was … because we didn't have quite the right algorithms.' However, he said, 'it turned out it was mainly a question of scale': early neural networks were simply not big enough. Hinton's former PhD student Ilya Suskever, co-founder of OpenAI (the start-up that developed ChatGPT), agrees:

> if you look at the history … for the largest time people thought all this neural networks, they can't do anything, but then you give them lots of compute [the capacity to perform very large numbers of computations] and suddenly they start to do things.

Making neural networks bigger wasn't easy. They learn by making predictions: what is the next word in this sentence? Is this image a cat? They then automatically adjust their parameters according to what the word actually is or whether a human being agrees that it's a cat. That learning requires large quantities of data – very big corpora of digitally-available text, lots of images labelled by human beings, etc. – and large amounts of 'compute'. Even as late as the early 2000s, AI faced limitations in those respects. A crucial aspect of the necessary computation is the multiplication of large matrices (arrays of numbers). If you do the component operations in those multiplications one after another on even a fast conventional computer system, it's

going to take you a long time, and you may find that you simply can't manage successfully to train a big neural network.

By about 2010, though, very big data sets were starting to become available. Particularly crucial was ImageNet, the establishment of which was led by Stanford University computer scientist Fei-Fei Li. It's a giant digital assemblage of millions of pictures, each labelled by a human being: Li and her colleagues recruited some 49,000 people via Amazon's Mechanical Turk platform, which enables the temporary hiring of large numbers of online gig workers. Around 2010, too, specialists in neural networks began to realise that they could do lots of matrix multiplications fast on graphics chips originally developed primarily for video games, especially by Nvidia.[1]

Those twin developments came together in 2012, in the single most important moment in launching AI on its trajectory of ever-increasing scale. Ilya Sutskever, Alex Krizhevsky (another of Hinton's students) and Hinton himself entered their neural-network system, AlexNet, into the annual ImageNet Challenge competition among automated image-recognition systems. Running on just two Nvidia graphics chips in Krizhevsky's bedroom, AlexNet won hands-down: its error rate was 40 percent lower than the best of its more conventional rivals.

The lesson that if you give neural networks 'lots of compute' by using graphics chips, then 'suddenly they start to do things' was quickly learned. From 2012 to around 2014, the number of graphics chips that a typical research project used was still modest: no more than about eight. But as neural networks moved from academia to

[1] Donald MacKenzie wrote about the importance to AI of the 'parallel programming' of Nvidia chips in the *LRB* of 20 November 2025.

tech companies, and from research to the development of AI models designed for practical use, exponential growth in scale set in. Jaime Sevilla and Edu Roldán of the research institute Epoch AI calculate that since the early 2010s, the amount of computation used in the training of state-of-the-art models has been increasing by between fourfold and fivefold annually. Fourfold annual growth implies 16-fold over two years, 64-fold over three years, and so on. That, in essence, is how you get from Krizhevsky's bedroom to Manhattan-scale data centres.

Nowhere is the imperative of ever-increasing scale more central than in OpenAI, set up in 2015. OpenAI's founders included Sutskever, Elon Musk and Sam Altman, who describes its ethos:

> When we started, yeah, the core beliefs were deep learning [neural networks with lots of 'layers'] works and it gets better with scale. ... And what took the like, the word that keeps coming to mind is like, religious, level of belief was that that wasn't going to stop.

The series of increasingly big language models developed by OpenAI gave a team of its researchers data that they were able to use to check that the firm's foundational 'like, religious' belief was empirically justified. Writing in January 2020, the team described how the accuracy of those models' predictions of the next word got better as their scale increased:

> Language modeling performance improves smoothly as we increase the model size [the number of parameters in a model], dataset size, and amount of compute used for training. ... We observe no signs of deviation from [these] trends at large values of compute, data, or model size.

Faith within OpenAI in these ‘scaling laws’ remains strong. As Sam Altman puts it in a February 2025 blog post:

> The intelligence of an AI model roughly equals the log of the resources used to train and run it. … It appears that you can spend arbitrary amounts of money and get continuous and predictable gains; the scaling laws that predict this are accurate over many orders of magnitude.

The ‘laws’ may of course break down – they are empirical generalisations, not laws of physics – but they are worth taking seriously because ‘arbitrary amounts of money’ are indeed being shelled out on the infrastructure of AI. In August, researchers for the investment bank Morgan Stanley estimated that $2.9 trillion is going to be spent globally on datacentres in 2025-28, while Citigroup has estimated total AI investment globally of $7.8 trillion in 2025-30. The latter figure is more than the entirety of likely US spending on defence over those years (its defence budget is currently around $1 trillion a year).

One little word, though, in what Altman writes should give us pause: ‘log’. A logarithmic function, at least of the kind that is relevant here, is characterised by diminishing returns (see the graph on the final page). The more resources you put in, the better the results, but the rate of improvement steadily diminishes. That’s not just an issue for specialists in AI, because the rest of us are also implicitly being taken on a ride along the logarithmic curve. The graphics chips and datacentres on which ‘arbitrary amounts’ are being spent require in aggregate giant quantities of electricity to power them. Some of this is coming from renewable sources, but much of it involves burning natural gas or sometimes even coal. Just one of the many new gas-fired power plants

that are being constructed in the US to meet the growing demands of data centres is on the site of an old coal-fired power station near Homer City, Pennsylvania. When it is up and running it will generate 4.4 gigawatts. For comparison, the peak winter electricity demand of the entirety of Scotland is only a little over 4 gigawatts.

One motivation for keeping going further along a diminishing-returns curve may well be the widespread sense that AI is a 'winner takes all' business, and so having the best models (by even a small margin) will bring disproportionate rewards, perhaps even de facto monopoly. Another motivation seems to be the hope that small improvements in performance will suddenly give birth to dramatic qualitative change: the emergence of 'artificial general intelligence' or maybe even 'superintelligence'.

The availability of humanly created data on which to train AI systems is, however, already a big constraint, argues Ilya Sutskever:

> compute is growing ... the data is not growing, because we have but one internet ... data is the fossil fuel of AI. It was like created somehow, and now we use it, and we've achieved peak data, and there'll be no more.

It's not a nuanced claim, but the specialists I've spoken to don't dismiss it out of hand. Soon, 'we'll be data constrained', says one, but he floats the prospect that AI systems themselves may be able to create reliable new data to replenish humanity's over-exploited reservoirs:

> What we don't know is, do we cross the threshold where AI produces novel training data, effectively? ... If we get to that threshold, then we are

> going to have superintelligence. We're very close. I don't think people understand how close we are, I don't think anyone wants to think about it.

However, another of the researchers to whom I have spoken, the natural language processing specialist Lonneke van der Plas, implicitly warns of the risk of training AI models on computer-generated data. Among the languages on which she works is Maltese, which has only around half a million native speakers. Much digitally available Maltese has been machine translated from other languages, often badly. In consequence, she says, if you simply go for scale in developing a model of Maltese 'you get a much worse system than if you carefully select the data' and exclude the reams of low-quality machine translation.

The huge amount of money being poured into AI infrastructure is also starting to worry the financial markets. Is AI a bubble, like the dot-com boom of the late 1990s? If it is, will the bubble burst, and if so when? It's impossible to be certain, but there's clearly a risk that many of the trillions of dollars that are being spent will turn out to have been wasted. The logarithmic curve would suggest a relatively benign scenario, in which the bubble slowly deflates as each successive new model turns out to be a little bit better than its predecessor, but the improvements become increasingly disappointing. But that's not how financial markets usually work. Fear of missing out keeps the bubble inflating, until something – perhaps in itself quite a small event – suddenly punctures it. Early in 2025, the successes of a large language model developed in a much less resource-intensive way by the Chinese firm DeepSeek temporarily threatened to do just that. Will something similar happen again over the coming months? If so, how bad will the economic damage be this time?

Fact checking links, etc

Hyperion data centre superimposed on Manhattan: https://www.threads.com/@zuck/post/DMF6tMAxkX8

Hinton quote: https://www.wired.com/story/ai-pioneer-explains-evolution-neural-networks/

Sutskever quote: https://www.youtube.com/watch?v=q_-kH-ybhFs

49,000 human beings recruited by Mechanical Turk: https://image-net.org/static_files/files/imagenet_ilsvrc2017_v1.0.pdf

Krizhevsky's bedroom: Cade Metz, *Genius Makers* (Penguin 2021), p. 95.

AlexNet error rate: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

'No more than around eight' chips: https://openai.com/index/ai-and-compute/

Sevilla and Roldán: https://epoch.ai/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year

Altman 'like, religious, level of belief' quote: https://x.com/tsarnick/status/1855001942919168207?s=61%20/%20religious%20belief

Kaplan et al., 'Scaling Laws for Neural Language Models': https://arxiv.org/pdf/2001.08361 (quote on pp. 3 and 17).

Altman blog post: https://blog.samaltman.com/three-observations

$2.9 trillion on datacentres: https://www.ft.com/content/efe1e350-62c6-4aa0-a833-f6da01265473

Citigroup $7.8 trillion estimate: https://www.ft.com/content/bd6ff6cd-3f27-49d7-930b-f3d83c0cb6c8

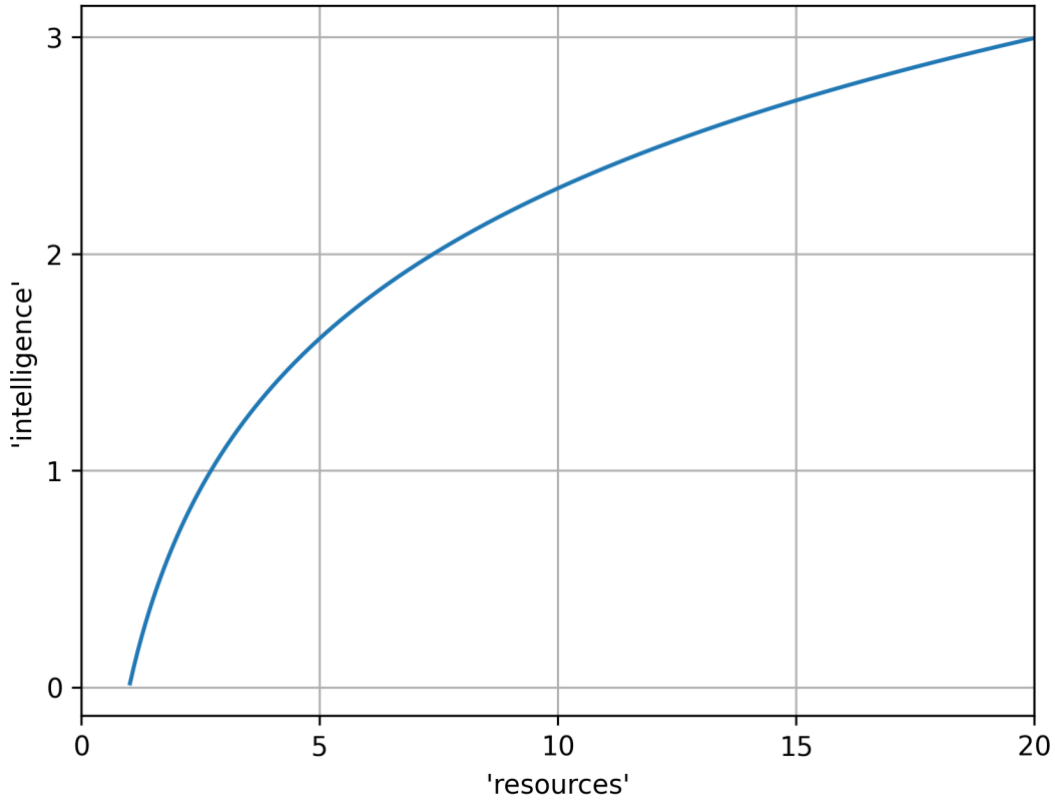
US defence budget: https://www.iiss.org/online-analysis/military-balance/2025/05/president-trumps-fy2026-defence-budget-continuing-priorities-new-missions/

Homer City 4.4 gigawatt plant: https://www.homercityredevelopment.com/project-overview

Scotland peak winter electricity demand: https://www.neso.energy/publications/electricity-ten-year-statement-etys/electricity-transmission-network-requirements/scottish-boundaries#:~:text=Scotland's%20current%20winter%20peak%20gross,reaches%20zero%20between%202035%20%2D%202040.

'Data is the fossil fuel of AI': https://www.youtube.com/watch?v=1yvBqasHLZs

Native speakers of Maltese: https://maltacentre.eu/maltese-language/



Part of the graph of the canonical ‘natural’ logarithm. Logarithmic curves such as this manifest diminishing returns: increments on the vertical access get smaller in each successive interval on the horizontal axis.