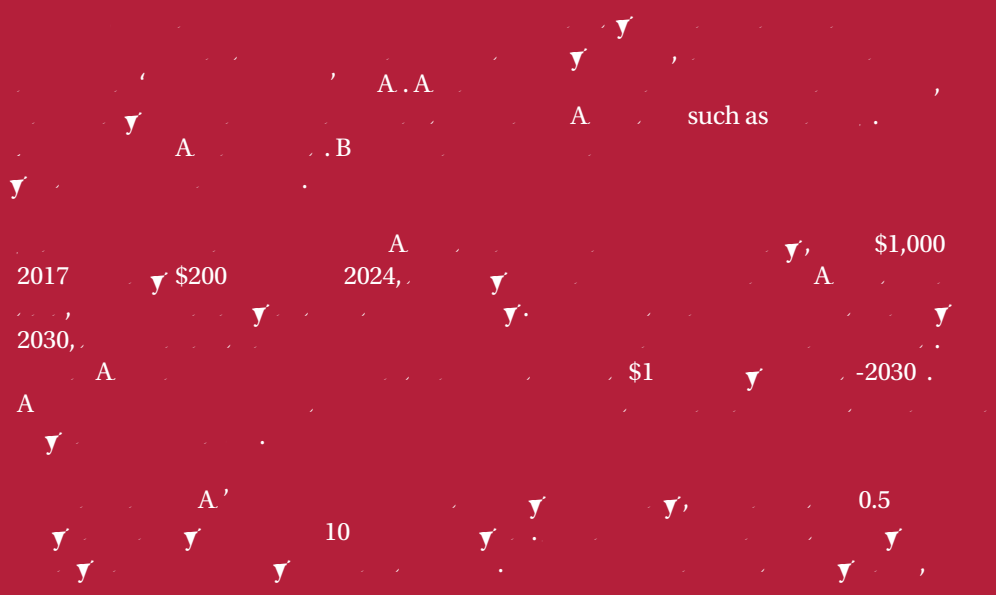
WORKING PAPER ISSUE 18/2024 09 OCTOBER 2024

THE TENSION BETWEEN EXPLODING AI INVESTMENT COSTS AND SLOW PRODUCTIVITY GROWTH

BERTIN MARTENS



1 Introduction

Optimism about artificial intelligence as a breakthrough technology with substantial economic impact has been growing since 2010. At that time, machine learning in neural networks started to show promise, in particular with the *"transformer"* deep-learning technology (Vaswani *et al*

2 Exponentially growing AI costs

GenAI models are statistical prediction models that need to be trained with large amounts of data (Agrawal *et al*, 2018). Older AI models required relatively small amounts of annotated data for training. Annotation identifies what the data is referring to. It is slow and costly because it is done by humans. They could only respond to questions related directly to the data on which they were trained. New generative AI models are much more flexible and require less annotation, but need much larger amounts of training data.

They also need more computing capacity to crunch the data during training.

largest frontier models at the end of 2023 – OpenAI's GPT4 and Google's Gemini Ultra – and extrapolating to 2030 would lead to an e

Aschenbrenner (2024, p 23) claimed even faster falling computing costs. But GenAI models increased demand for computing capacity by about twelve orders of magnitude between 2010 and 2024 (Aschenbrenner, 2024, p 21). Figure 1 suggests an expansion of eight orders of magnitude between 2016 and 2023, from 1000 to 100 billion petaflops per GenAI model. Clearly, the quantity e□ect dominates the price e□ect.

The current wave of GenAI models is trained on human-

3 The benefits of AI: productivity growth

The main economic benefit expected from AI is productivity gains. Humans will be able to complete tasks faster and more e□ciently. But what is known about these productivity gains?

Brynjolfsson *et al* (2020) made a general observation on technology-induced productivity gains. They argued that embedding digital technology in firms requires costly investment in all kinds of complementary tasks and activities. That slows down the productivity uptake of new technologies and initially drags down productivity before it rises fast afterwards when the benefits mature. They found evidence of this productivity J-curve e□ect in computer software and suggested that this may also be the case for AI. Brynjolfsson *et al* (2020) recognised the productivity potential of AI but underscored that the AI roll-out across the economy can be expected to be much slower than the development of AI models.

The J-curve roll-out e□ect is spread out over time, across the entire economy and across a wide range of GenAI models, not only the most powerful models at the technology frontier. Much of the roll-out across the economy will come from AI models below the technology frontier, including smaller models that can be trained with far less computing power than large foundational models, compressed models that are derived from large frontier models but redesigned to run at far lower computing costs[8] (Grootendorst, 2024) and specialised models designed for specific tasks. Developers of large foundation models are building ecosystems of satellite models around core large models. For example, OpenAI set up a ChatGPT applications store that contains millions of specific applications of ChatGPT. OpenAI also made available a much more compact version of its leading GPT4 model, ChatGPT4o-mini, designed to run on laptops and smartphones for all kinds of daily uses, such as children doing their homework or parents planning holidays. This branching out of large AI models into many derived models and applications that build on top of the foundation GenAI model will help to amortise the huge fixed costs of foundation AI models across a wide range of applications.

The cognitive returns to AI constitute the primary driver to boost productivity growth: completion of tasks by humans will be done more cheaply by machines. However, Acemoglu (2024) saw only limited prospects for human-machine substitution and productivity growth. He estimated that an AI-driven productivity increase will not exceed 0.5 percent in the next decade. By contrast, Goldman-Sachs economists put that estimate at nearly 10 percent (Nathan *et al*, 2024). The di□erence between the two arises from Acemoglu's very conservative estimates of the share of human tasks that will be a□ected by AI, cost savings and expansion of new tasks, and more capital deepening in the economy. The substitution approach has very little to say about the emergence of cognitive tasks that humans cannot carry out because of cognitive limitations on human brainpower, or on the automated automation of tasks.

[8] See Maarten Grootendorst, 'A Visual Guide to Quantization', 22 July 2024, https://www.maartengrootendorst.com/blog/quantization/.

for example, health, environment, transport and combating climate change[10]. For example, OpenAI, one of the market leaders in AI, was originally established as a non-profit organization with the 'AI for good' purpose in mind. OpenAI is now a private

energy. Mankind has fortunately not reached that point yet. But the rapid development of AI models is moving in that direction. Investment in AI computing power will continue and mankind will find the means to pursue that goal, possibly beyond economic rationality.

It is obvious that single- or even double-digit productivity growth scenarios are unable to keep pace with exponentially growing AI investment costs. At best, the point of reckoning might be postponed. The qu ƀvith (i6e)5 (v3 (e)-5 (c)-6 (d)]TJ0 Tc9 (a17.45 0 Td(Tw 0.17(– Td(56w 1.18 0 Td( )Tj0.004 Tc 0.001 Tw 0.18 0o

References

Agrawal, A., A. Goldfarb and J. Gans (2018) *Prediction Machines: The Simple Economics of Artificial Intelligence*, Harvard Business Review Press

Alderucci, D., S. Baviskar, L. Branstetter, N. Goldschlag, E. Hovy, A. Runge, P. Tambe and N. Zolas (2024) ‘tta(s)6 5 ( )]10.

IMF (2024) *World Economic Outlook*, April 2024, International Monetary Fund, available at https://www.imf.org/en/Publications/WEO/weo-database/2024/April

Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger … D. Hassibis (2021) 'Highly accurate protein structure prediction with AlphaFold', *Nature* 596: 583–589

Kaplan, J., S. McCandlish, T. Brown, B. Chess, R. Child, S. Gray … D. Amodei (2020) 'Scaling Laws for Neural Language Models', mimeo, available at https://arxiv.org/abs/2001.08361

Lu, C., C. Lu, R.T. Lange, J. Foerster, J. Clune and D. Ha (2024) 'The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery', mimeo

Bruegel, Rue de la Charité 33, B-1210 Brussels
(+32) 2 227 4210
info@bruegel.org
www.bruegel.org