WORKING PAPER | ISSUE 18/2024 | 09 OCTOBER 2024

THE TENSION BETWEEN EXPLODING AI INVESTMENT COSTS AND SLOW PRODUCTIVITY GROWTH

BERTIN MARTENS

This working paper explores the tension between rapidly increasing artificial intelligence investment costs and the slower pace of productivity growth, raising concerns about a potential 'economic winter' for AI. AI has shown significant technological progress, particularly with machine learning and generative AI models such as ChatGPT. Investment in AI has surged. But there are concerns about whether these investments will yield proportional returns.

Training costs for a single frontier AI model are increasing exponentially, from $1,000 in 2017 to nearly $200 million in 2024, driven by constant returns to scale in AI model training data, compute capacity and model complexity. Costs could reach billions of dollars by 2030, despite a rapid fall in unit costs per computing operation over the same period. Global AI infrastructure costs in hardware could exceed $1 trillion by the mid-2030s. Amortizing these huge fixed costs requires business models that can be rolled out across a very large user market.

Estimates about AI's contribution to productivity growth vary, from a modest 0.5 percent per year to a very optimistic 10 percent per year. Research shows that productivity usually catches up slowly compared to costs. Without significant productivity gains, the current investment cost trajectory is unsustainable. We estimate that 3 percent annual productivity growth across advanced economies would be required to sustain AI model cost extrapolations to 2030. In an optimistic scenario, productivity increases would result in accelerated economic growth.

Bertin Martens (bertin.martens@bruegel.org) is a Senior Fellow at Bruegel

bruegel

Recommended citation:
Martens, B. (2024) 'The tension between exploding AI investment costs and slow productivity growth,' *Working Paper* 18/2024, Bruegel

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*, 2017) that led to the current generation of large language models or generative AI (GenAI) models, including ChatGPT[1]. Since then, AI has been riding a wave of astonishing technological progress and massive investment. It has propelled stock market valuations of big-tech companies to dizzying heights.

But analysts worry that AI investment spending will exceed returns[2], both for AI model developers and users. Companies can only afford the AI expenditure if it increases their revenues, but nevertheless, they are continuing to double down on AI capital expenditure, arguing that they cannot afford to miss the opportunities that AI will bring. Stock markets seem to reward them for doing so. How long can this tension between spending and revenue continue? If the current pace of exponential growth in the size of GenAI models continues, AI investment costs would reach macro-economic significance within the next five years and become totally unaffordable within a decade. How much GenAI-driven productivity growth would be needed to keep pace with that cost trajectory?

This working paper looks at the drivers of the currently fast-growing AI costs and the economic conditions under which continued cost growth may or may not be sustainable. It takes inspiration from the analysis of dual-speed economies by Nordhaus (2021) to explore some medium-term economic scenarios for AI. Will a rapidly growing share of AI expenditure in GDP accelerate growth, or will AI absorb too much capital without a concomitant rise in productivity, and lead to an economic 'AI winter'? AI has been through 'winter' periods before, when investment dried up because of lack of technological progress[3]. This time, investment may dry up because returns do not justify the costs. Are there alternative AI trajectories that would avoid this scenario?

Section 2 provides some empirical evidence on exponentially growing GenAI investment costs, cost components and the scaling law that drives exponential growth. Section 3 takes an economic perspective and compares that cost explosion with the benefits side, in the form of GenAI-induced productivity growth. Section 4 brings the cost and benefits side together and shows that they are on an unsustainable trajectory. Section 5 goes beyond economics and explores some non-economic drivers of AI investment. Section 6 concludes.

[1] Large language models and generative AI models respond to prompts to produce, respectively, new text and new images or other audiovisual material, based on 'learning' from very large datasets.

[2] Skye Jacobs, 'Big Tech needs to generate $600 billion in annual revenue to justify AI hardware expenditure', *TechSpot*, 7 July 2024, https://www.techspot.com/news/103699-big-tech-needs-generate-600-billion-annual-revenue.html; Hannah Murphy, 'Meta's revenue growth reassures investors as Zuckerberg plots AI spree', *Financial Times*, 31 July 2024, https://www.ft.com/content/edbe2580-0b64-4339-9be2-4b4fee46211b; Camilla Hodgson, 'Microsoft's slower cloud growth fails to impress Wall Street', *Financial Times*, 1 August 2024, https://www.ft.com/content/cc781bd9-315a-4e9c-9730-393ca14fbcbb.

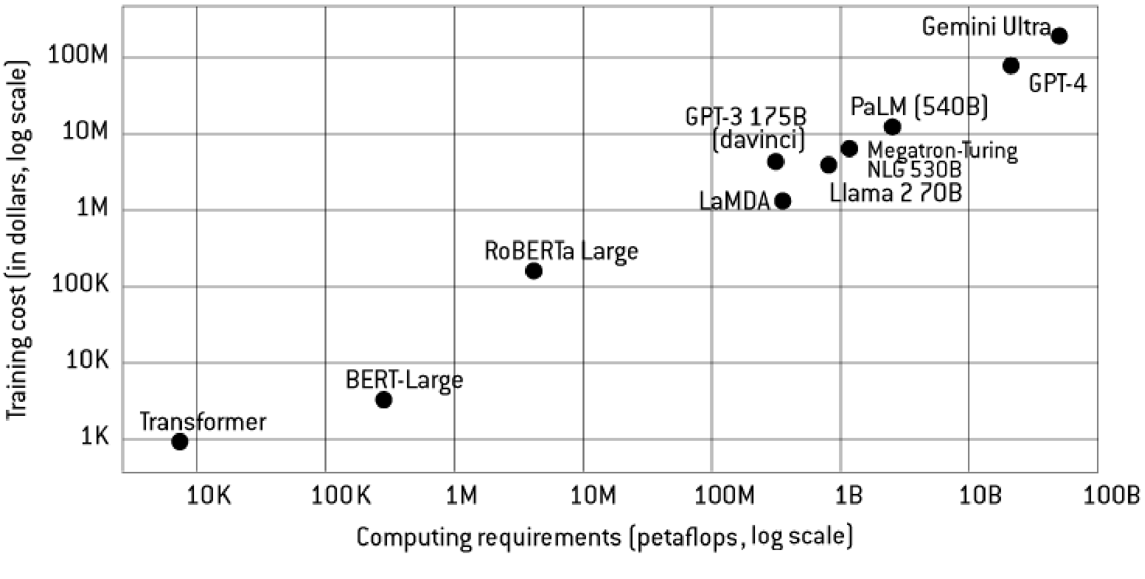
[3] See for example https://en.wikipedia.org/wiki/AI_winter.

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. Figure 1 illustrates the exponential growth in computing requirements, starting with Google's transformer model (Vishwani *et al*, 2017), which laid the foundation for generative AI technology. Computing requirements, expressed in petaflops[4], increase by an order of magnitude (x10) for each new generation of models, from less than 10,000 petaflops for the first transformer model in 2016 to more than 100 billion petaflops for recent models (2023). Computing costs have been rising proportionally to computing requirements (Figure 1, y axis), from around $1000 for the first transformer model in 2017, to Google's Gemini Ultra model in 2023 that cost $120 million to train. As of late 2024, training costs for top-ranking GenAI models are nearer to $200 million.

Figure 1: Training costs and computing needs for selected AI models



Source: Bruegel based on Epoch (2023) and AI Index Report 2024, p 65-66.

Cottier *et al* (2024) estimated that GenAI model training costs increased exponentially by a factor of 2.4 to 2.6 per year, or around 240 percent per year, from 2016 to 2023. Starting from the cost of the

[4] FLOP stands for 'FLOating-Point operation'. A floating-point operation is a single arithmetic operation involving floating-point numbers, such as addition, subtraction, multiplication or division. Flops are a measure of the computational power of a computer. 1 petaflop = 10^{15} flops per second.

largest frontier models at the end of 2023 – OpenAI's GPT4 and Google's Gemini Ultra – and extrapolating to 2030 would lead to an estimated training cost for a single GenAI model of $60 billion. Currently, there are dozens of frontier GenAI models, and new ones are coming out every month. Further extrapolation to 2035 would produce an almost unimaginable figure of nearly $6 trillion, nearly half of EU GDP. That does not look very plausible.

On top of training costs, AI requires investment in hardware and infrastructure in the form of data centres. Cottier *et al* (2024) estimated the cost of computing infrastructure at ten times the cost of model training. That infrastructure can be used to train several models. The same authors estimated the hardware amortisation rate at 140 percent/year, or 100 percent depreciation in 8.5 months[5]. By that time, a new generation of AI computing chips will have arrived with superior performance. Infrastructure costs for GPT4 by end 2023 may have been as high as $800 million (Cottier *et al*, 2024). Extrapolation could push that figure up to $500 billion by 2030 and many trillions of dollars by the mid-2030s[6]. And that infrastructure cost could be replicated across at least half a dozen hyper-scaling big tech firms. When infrastructure is written off for model training, it can still be used for other purposes, including for 'inference' or daily running of GenAI models for business use to reply to user queries. Running costs come on top of training costs.

Cottier *et al* (2024) broke down model training costs into five components: staff, AI chips, other server costs, network interconnection costs and energy consumption. In accounting for hardware costs, they assumed that the AI modeller owns the hardware that is amortised over the duration of the model training period. Alternatively, modellers can pay market rental rates for cloud infrastructure during the training period. Cost estimates cover the entire development cycle of an AI model, from pre-training experiments needed to design new model features and components, to model calibration and the final model training run.

Figure 2 presents an estimated break-down of these five components for some recent AI models. Staff is often the largest cost component. Skilled AI staff are scarce. AI companies compete for each other's employees. Staff are usually paid a salary plus equity benefits, which are hard to estimate, especially for AI start-ups with volatile stock values. Staff costs include data collection and cleaning costs, often carried out by large teams of lower-skilled staff.

Specialised AI accelerator chips are the second most expensive component. The market for AI chips is very monopolistic, with a single market leader (Nvidia). Markets for hyper-scaled cloud infrastructure are also tightly controlled by half a dozen players. Other server and cluster interconnection costs account for the remaining model training costs. Most of these inputs are sold in markets that

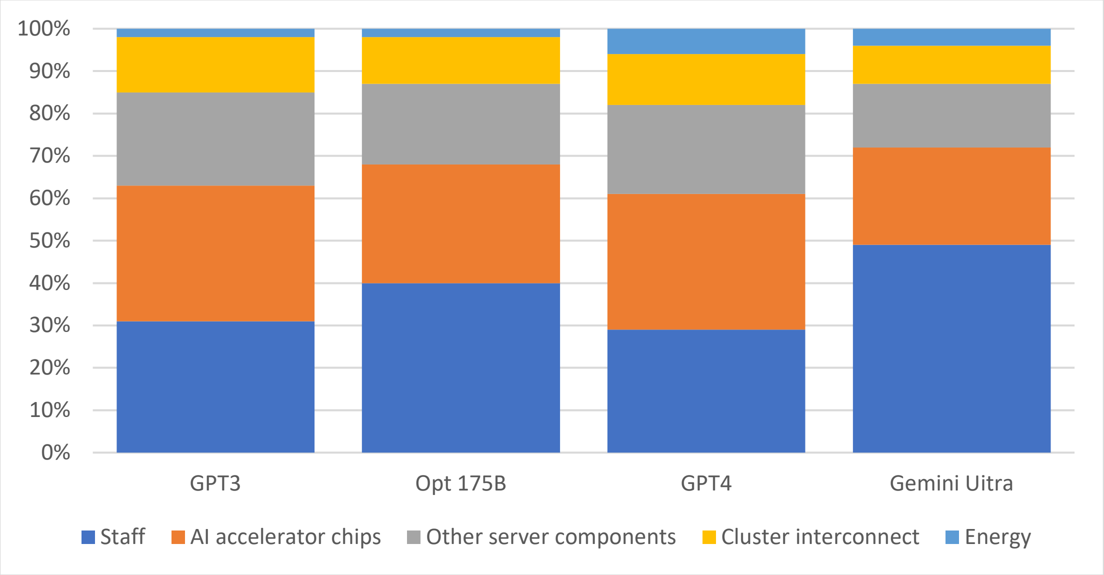
[5] This depreciation rate is based on the estimated growth rate in processor price-performance (Hobbhahn *et al*, 2023). This depreciation rate is for AI processor chips only. Other hardware parts of the cloud computing infrastructure may have longer economic lifetimes.

[6] With that perspective in mind, Sam Altman's search for investors to finance a trillion dollar AI computing centre suddenly sounds more plausible, though not necessarily more realistic. See Keach Hagey, 'Sam Altman Seeks Trillions of Dollars to Reshape Business of Chips and AI', *Wall Street Journal*, 8 February 2024, https://www.wsj.com/tech/ai/sam-altman-seeks-trillions-of-dollars-to-reshape-business-of-chips-and-ai-89ab3db0.

experience supply bottlenecks that drive up prices (Martens, 2024). Electricity is only 4 percent to 6 percent of the total. Power supply for server farms may come from local suppliers. Data centres might also set up their own power plants. Note that electricity costs are estimated for AI developers in the United States. In the EU, electricity prices are several times higher than in the US.

Training costs may vary considerably between models. Google's Gemini Ultra uses Google's own Tensor Processing Units (TPU). That avoids paying mark-ups to external chip suppliers. Moreover, TPUs are an order of magnitude more efficient in computing terms because they operate at lower eight-bit precision levels compared to the 32-bit standard (Hobbhahn *et al*, 2023). Other AI computational performance factors affect cost outcomes, including chip-to-chip communication bandwidth that increases with new technologies.

Figure 2: Breakdown of costs for AI model training and experimentation



Source: Bruegel based on Cottier *et al* (2024), p 9.

The main driver of the AI model training cost explosion is the so-called 'scaling law' for GenAI models. Kaplan *et al* (2020) found that improvements in the cognitive performance of GenAI models are subject to constant returns to scale in complementary inputs, including the number of model parameters, the volume of training data and the available computing capacity. For optimal model performance, all three factors must be scaled up simultaneously. Diminishing returns occur when one of these three factors is not scaled accordingly. The volume of training data only needs to be increased by 0.74 times the increase in the number of parameters, indicating that larger models are more data efficient than smaller models. Kaplan *et al* (2020) expected constant returns to scale to continue to hold for the foreseeable future. The red line in Figure 1 illustrates that point.

The cost explosion is happening despite a rapid decline in computing costs. Hobbhahn *et al* (2023) estimated that the cost of computations halved every 2.1 to 2.5 years from 2016 to 2022.

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 effect dominates the price effect.

Vast increases in the scale and complexity of AI models have contributed to substantial improvements in cognitive performance in the last couple of years, as measured by the best human performance benchmarks on many tasks. The most advanced models are now performing close to or even above human benchmarks on tasks including image recognition, language understanding and mathematics (Maslej *et al*, 2024). New performance benchmark tests that exceed human performance will be needed. Still, there are basic tasks for which human intelligence is still far superior to AI. For example, 3D vision and visual sorting of objects, tasks that can easily be accomplished by children, are still problematic for AI, though improvement is within reach with further scaling. A general AI model that outperforms humans in every aspect, is still far away, if it is even possible.

It may be possible to achieve increasing returns to scale and to increase the efficiency of input use (data, computing capacity) in GenAI models. Kaplan *et al* (2020) calibrated their scaling law on the current generation of transformer AI models that started in 2017. New ways of using these models may affect the scaling law, in particular for data.

Algorithmic efficiency gains can reduce the computational capacity required to train a model. Erdil and Besiroglu (2022) estimated that the amounts of computations required for training the same AI model decreased by about 500 percent per year between 2012 and 2021 because of increases in the efficiency of AI training algorithms. These gains enable AI models to train on smaller datasets, with lower computing capacity, to achieve the same outcomes. For example, OpenAI's GPT-4o is much smaller than ChatGPT-4, can be implemented on less powerful computers and still achieves comparable outcomes. Quantum computing is a promising technology to reduce the cost of computation by many orders of magnitude. However, that depends on if and when quantum computing capacity can be scaled up sufficiently to accommodate large AI models. That is not yet within reach.

Another potential source of increasing returns is what Aschenbrenner (2024) called the *"unhobbling"* of AI models – finding new ways in which AI models deal with problems. Each model has its own approach to problem solving and sticks to that approach. But there may be other and more efficient ways to solve problems. Human feedback through prompting and 'chain-of-thought' reasoning[7] may give models human guidance on how to approach problems. Combining several AI models in a collaborative division-of-labour may also make models more efficient. AI engineers are experimenting with expanding the context or explaining the problem setting. The first GenAI could only handle 2000 words of context; the latest models can handle more than 1 million words.

[7] See IBM, 'What is chain of thoughts (CoT)?' undated, https://www.ibm.com/topics/chain-of-thoughts.

The current wave of GenAI models is trained on human-generated data. They learn to imitate human behaviour. However, GenAI models are now being used in configurations that depart from this human imitation scenario and facilitate learning and knowledge generation without human behavioural data, or with a small patch of human data as examples to initiate a learning process. For example, Google DeepMind developed the AlphaFold AI model (Jumper *et al*, 2021) to predict possible protein structures. The model learned the basics of protein structures from about 100,000 human-discovered protein structures and then set out to discover billions of additional protein structures.

Two or more AI models can also be put in an adversarial configuration in which they compete to discover better responses to challenges. This is how DeepMind's AlphaGo model discovered new strategies to play the Go game, superior to known human strategies, starting only from a set of basic rules for the game. Adversarial configurations are widely applied in strategic settings, such as military and security applications of AI models, to discover more efficient strategies that have never been played by humans, or would have been impossible for humans because they require very fast data processing and response times that humans are unable to muster (Scharre, 2023).

A move away from the imitation of human behavioural data will become increasingly necessary because the rapidly expanding data requirements of GenAI models are hitting the so-called data wall, the maximum amount of human-generated data available (Maslej *et al*, 2024). Frontier AI models currently need many trillions of tokens or training data points. There is not enough text data on the internet and in other text archives to meet the requirements of the largest models. Stringent application of copyright law and opt-outs may further reduce the volume and increase the cost of available training data (Martens, 2024b).

AI model developers are trying out various methods to overcome this constraint, for instance converting speech into text or using synthetically generated data. Some claim that synthetic data reduces model quality and may result in the complete collapse of AI models (Shumailov *et al*, 2024). Others are more optimistic (Ben Allal *et al*, 2024). Generating new data and knowledge in AI training settings that reach beyond what humans can produce is thus a necessary condition for continued growth in AI. That will inevitably steer the evolution of AI models away from transformer-type GenAI models, and thus from the scaling law that these models are subject to.

Aschenbrener (2024) argued that an important step in generating non-human data and knowledge will be reached when AI can write its own code and solve its own research questions. Lu *et al* (2024) claimed to have achieved this task, at least for AI research (this claim at time of writing has not had wider confirmation or rejection). Nordhaus (2021) reformulated this is a more general way: AI should be able to take over and automate task automation, rather than leaving it to humans to invent task-automation solutions.

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 efficiently. 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 effect 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 effect 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 difference between the two arises from Acemoglu's very conservative estimates of the share of human tasks that will be affected 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/.

The empirical evidence on AI productivity effects is not yet very insightful. Several studies found positive productivity impacts of earlier AI technologies that pre-dated the arrival of GenAI (Alderucci *et al*, 2024; Da Silva *et al*, 2024; Czarnitski *et al*, 2022). There have been so far very few studies on the productivity effects of the current wave of GenAI models. Noy and Zhang (2023) found that ChatGPT raises productivity and quality in writing tasks by 0.8 and 0.4 standard deviations respectively. Brynjolfsson *et al* (2023) found a 10 percent productivity increase when call-centre operators are assisted by GenAI models. There are too few studies to generalise the findings. However, the number of firms that subscribe to ChatGPT and other GenAI models for professional use by their employees is growing very rapidly. An average professional business subscription currently amounts to less than €60/month/employee. That low price ensures wide penetration in business markets. Widespread use is a good indication that businesses expect significant returns from their investments in AI subscriptions.

4 Bringing the cost and benefit sides together

At the current rate of exponential cost growth, AI infrastructure investment and modelling costs will reach macro-economic significance over the next decade. The International Monetary Fund World Economic Outlook 2024 (IMF, 2024) estimates that World GDP will reach $139 trillion by 2029, with 27 percent of GDP or $37 trillion going to investment. The WEO forecast does not extend to 2030 or 2035. But a simple extrapolation of the 2025-2029 3.1 percent real growth rate to 2035 would bring world GDP to around $143 trillion in 2030 and $156 trillion in 2035, of which about $36 trillion and $40 trillion respectively would go to investment. Assuming that there would still be competition between at least five very large firms with their own state-of-the-art AI hardware infrastructure, AI investment could reach $2.5 trillion in 2030 and $250 trillion in 2035. That would amount to a whopping 7 percent of world GDP in 2030 and an impossible 160 percent of GDP in 2035. Clearly, these growth rates are not sustainable and something will have to give.

How can the rapidly escalating costs of AI be matched with the much slower emergence of AI productivity benefits?

Nordhaus (2021) suggested following Baumol's (1967) demand-side perspective on a two-speed economy, in which a rapidly advancing technology is subject to declining prices that are not matched by growth in demand. As a result, the share of that technology in GDP declines and the increasing relative cost of traditional technologies slows down GDP growth. Baumol called this *"cost disease"*. The opposite can also happen: rapidly declining prices for new technologies can trigger fast growth in demand, increase their share of GDP and accelerate GDP growth. However, Nordhaus (2021) showed that even for information technology sectors that experienced rapid price declines over the last two decades, including phones, computers and telecommunication services, there is no conclusive evidence either for rising or declining shares of these sectors in GDP.

The issue can also be examined from a supply-side perspective, comparing a quasi-constant or slow-growing volume of human and other types of capital with a rapidly growing volume of information capital. The volume of information capital – in computers, software, data or AI – in GDP will rise if the volume of information capital expenditures rises faster than the relative price of information capital declines, compared to other capital. Nordhaus (2021) called this *"growth euphoria"* on the supply side – the reverse of Baumol's (1967) *"cost disease"*: the quantity effect dominates the price effect and boosts GDP growth, to the point where a growth acceleration might emerge, not only through Keynesian multiplier and accelerator effects, but also because the increasing share of investment in GDP further accelerates GDP growth. The point might be reached where investment dominates economic growth and triggers a significant growth acceleration – an economic *"singularity"* in Nordhaus's words. Using US data from the past decades, he found that a rising share of capital in GDP and a rising share of information capital in total capital pass the singularity test, but only just. Their growth trajectories are so slow that any singularity or growth acceleration event is at least a century away.

Nordhaus (2021) used relatively old data for his empirical test, dating back to before the arrival of GenAI. With exponential growth in the cost of AI models, the cost of information capital in total capital and in GDP is rising much faster than implied by Nordhaus's historical data for digital technologies. Would that bring the singularity horizon nearer?

A simple back-of-the-envelope productivity growth calculation can clarify this. Assume that digital firms represent 10 percent of GDP and jointly invest $1 trillion in AI models, a figure that could easily be reached in the next few years. Assume also that this triggers a significant 3 percent productivity increase in 90 percent of the economy, generating a 2.7 percent increase in total GDP. Note that 3 percent would already be quite an improvement compared to the 1.5 percent or so current rate. The GDP or market scale required to reach break-even for these AI investment costs would be $37 trillion if all productivity gains accrue to the digital firms that invest in AI[9]. More realistically, if only 50 percent of the gains accrue to digital firms, and the rest to AI users, the required GDP would be $72 trillion. According to the IMF (2024), GDP for the advanced economies amounts to $75 trillion. A 1 trillion AI investment is thus a credible choke-point, unless the AI investment were to continue to expand and/or trigger even higher productivity growth, above 3 percent, to achieve a growth 'singularity'.

5 Non-economic drivers of AI investment

The reasoning above is confined to purely economic considerations. But other factors drive AI investment, beyond economic rationality. Human emotions and fears are an important push factor in AI investment.

People are motivated by the 'altruism factor': a large part of the AI scientific community is driving 'AI for good', the application of AI technology to good causes that will improve human wellbeing in terms of,

[9] Productivity gains of 2.7 percent on a GDP of $37 trillion amount to $1 trillion in economic value gains. That covers the AI investment cost.

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 for-profit company because AI has become big business that requires very large investments. An important part of that altruism movement is focused on ensuring alignment of AI model responses with human values. Developing general AI, on par or exceeding the capacities of human intelligence, is considered by some as an important contribution to the well-being of mankind[11], but by others as a major threat to mankind (McLean *et al*, 2023).

Firms are motivated by the 'chicken game factor'. None of the big tech companies involved in AI development is willing to reduce AI investment, fearing they will miss out on a big opportunity. Investors subscribe to that view and downgrade the stocks of companies that show any sign of slowing down on AI investment, even when profits are under pressure[12]. Of course, investors know that new technologies are subject to hype-cycles of exuberance and crashes. But they will want to ride the cycle until the very last moment before the crash.

Countries meanwhile are motivated by the 'China factor': AI is perceived as the most important frontier in military and security technology (Scharre, 2023). Mastery of the most advanced AI technologies will determine the outcome of future conflicts. The US and China are locked in an AI arms race they cannot abandon, whatever the costs (Scharre, 2023). As was often the case in the past, arms races are a major driver of technological development. General AI (Goertzel, 2014), the holy grail of AI, is seen as the ultimate military technology (McLean et al, 2023). AI as a non-rival digital technology can overcome any hardware and manpower constraints in conflict situations, provided unlimited computing capacity and data are available. This argument could be more generally reformulated as a national security factor that becomes increasingly important in a rapidly fragmenting and polarising geopolitical world. It also holds for the EU and its attempts to boost EU AI investment. Should the EU remain dependent on US big tech firms in terms of computing capacity and AI technology?

6 How will this end?

Nordhaus (2021) reaffirmed that AI is the first human-invented technology that replaces not only human labour but also human reasoning and knowledge accumulation – a production factor on which mankind has had a natural monopoly until very recently. Since there is no limit to knowledge growth, Dyson's (1960) observation should perhaps be heeded: highly intelligent societies are necessarily computing-intensive. AI machines are needed to overcome human cognitive capacity constraints. Since computing requires energy, Dyson predicted that super-intelligent societies will want to capture all the energy available to feed their computers. They would literally burn the planet to harvest that

[10] See for example the 'AI for Good' summit meetings, organised by the United Nations, and the very broad policy agenda that it promotes: https://aiforgood.itu.int/.

[11] OpenAI's mission statement refers to this. See OpenAI, 'Planning for AGI and beyond', 24 February 2023, https://openai.com/index/planning-for-agi-and-beyond/.

[12] Intel is a recent example. See Anton Goncalves, 'Intel's weak position in AI chip market leads to mass layoff', *TechTarget*, 2 August 2024, https://www.techtarget.com/searchdatacenter/news/366599761/Intels-weak-position-in-AI-chip-market-leads-to-mass-layoff.

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 quantity effect – increased demand for computing power for AI – will continue to overwhelm the price effect of rapidly falling computing costs. That brings the economy closer to Nordhaus's (2021) *"growth euphoria"* scenario in which the share of AI investments in GDP will increase – until economic, political and social pressures slow it down. A pessimistic new 'AI winter' scenario in which AI investments freeze for economic reasons is certainly possible. An optimistic scenario with accelerating productivity and GDP growth rates looks less likely at this point, unless Nordhaus's (2021) ultimate AI scenario becomes a reality: AI automating the automation of production processes. Researchers have already taken steps in that direction (Lu *et al*, 2024).

Less dramatic intermediate scenarios look more likely. Technological change in computer hardware, software and AI model architecture may bend the GenAI scaling law of Kaplan *et al* (2020) towards increasing returns to scale. That could fend off, or at least postpone, the economic-freeze scenario. Moreover, the scaling law applies to GenAI models only. They may be replaced by other types of AI model that will be less dependent on human-generated data for training. They may generate their own learning and knowledge.

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) 'Quantifying the Impact of AI on Productivity and Labor Demand: Evidence from U.S. Census Microdata', mimeo, available at https://www.aeaweb.org/conference/2020/preliminary/paper/Tz2HdRna

Aschenbrenner, L. (2024) 'Situational awareness, the Decade Ahead', mimeo, available at https://situational-awareness.ai/

Baumol, W. (1967) 'Macroeconomics of Unbalanced Growth: The Anatomy of Urban Crisis', *The American Economic Review* 57(3): 415-426

Ben Allal, L., A. Lozhkov and D. Van Strien (2024) 'Cosmopedia: how to create large-scale synthetic data for pre-training', Blog Post, *Hugging Face*, 20 March, available at https://huggingface.co/blog/cosmopedia

Brynjolfsson, E., D. Rock and C. Syverson (2020) 'The productivity J-cuve: how intangibles complement general purpose technologies', *NBER Working Paper* 25148, National Bureau of Economic Research

Brynjolfsson, E., D. Li and L. Raymond (2023) 'Generative AI at work', *NBER Working Paper* 31161, National Bureau of Economic Research

Cottier, B., R. Rahman, L. Fattorini, N. Maslej and D. Owen (2024) 'The rising costs of training frontier AI models', mimeo, available at https://arxiv.org/abs/2405.21015

Czarnitski, D., G. Fernandez and C. Rammer (2022) 'AI and firm-level productivity', *ZEW Discussion Paper* 22-05, ZEW – Leibniz Centre for European Economic Research

Da Silva L., A. Rincon-Aznar and F. Venturini (2024) 'Productivity performance, distance to frontier and AI innovation: Firm-level evidence from Europe', mimeo, available at https://dx.doi.org/10.2139/ssrn.4728017

Dyson, F. (1960) 'The search for artificial interstellar infrared radiation', *Science* 131(3414): 1667-1668

Erdil, E. and T. Besiroglu (2022) 'Algorithmic progress in computer vision', mimeo, available at https://arxiv.org/abs/2212.05153

Hobbhahn, M., L. Heim and G. Aydos (2023) 'Trends in Machine Learning Hardware', *EpochAI*, 9 November, available at https://epochai.org/blog/trends-in-machine-learning-hardware

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, available at https://arxiv.org/abs/2408.06292

Martens, B. (2024) 'Economic arguments in favour of reducing copyright protection for generative AI inputs and outputs', *Working Paper* 09/2024, Bruegel

Martens, B. (2024) 'Why artificial intelligence is creating fundamental challenges for competition policy', *Policy Brief* 16/2024, Bruegel

Maslej, N., L. Fattorini, R. Perrault, V. Parli, A. Reuel, E. Brynjolfsson … J. Clark (2024) *The AI Index 2024 Annual Report*, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University

McLean, S., G. Read, J. Thompson, C. Baber, N. Stanton and P. Salmon (2023) 'The risks associated with Artificial General Intelligence: A systematic review', *Journal of Experimental and Theoretical AI* 35(4): 1-17

Nathan, A., J. Grimberg and A. Rhodes (2024) 'GenAI: too much spend, too little benefit?' *Global Macro Research* Issue 129, Goldman Sachs

Nordhaus W. (2021) 'Are We Approaching an Economic Singularity? Information Technology and the Future of Economic Growth', *American Economic Journal: Macroeconomics* 13(1): 299–332

Noy, S. and W. Zhang (2023) 'Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence', *Science* 381(6654): 187-192

Scharre P. (2023) *Army of None: Autonomous Weapons and the Future of War*, W W Norton

Shumailov, I., Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson and Y. Gal (2024) 'AI models collapse when trained on recursively generated data', *Nature* 631: 755-759

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser and I. Polosukhin (2017) 'Attention Is All You Need', mimeo, available at https://arxiv.org/abs/1706.03762

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