International Center for Future Generations ICFG

Advanced AI: Technical State of Play

Compute and Capabilities

Author: Daan Juijn

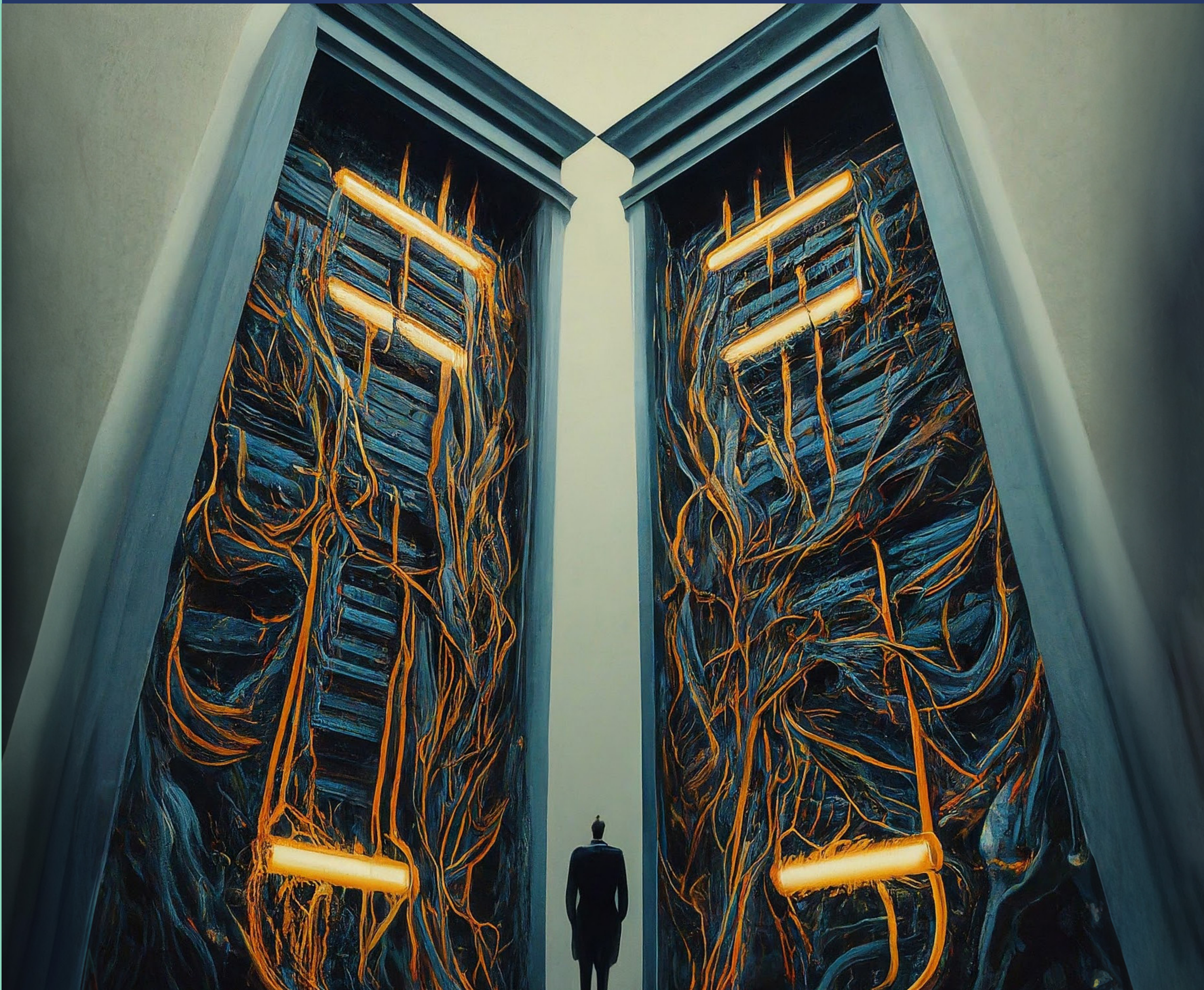


Table of Contents

Executive Summary

Advanced AI systems are evolving rapidly, revolutionizing industries and societies with capabilities like sophisticated language processing, text-to-speech conversion, and lifelike imaging. This surge is primarily fueled by exponential increases in computing power used to train advanced AI systems: the largest AI models of today are trained using billions of times more mathematical operations than previous state-of-the-art systems required back in 2010.

Over the next five years, compute, alongside algorithmic advancements, is expected to continue to drive the exponential progress in advanced AI. With AI companies increasing their compute budgets by more than fourfold annually, breakthrough AI systems—such as highly proficient autonomous agents—could soon emerge. These systems have the potential to rejuvenate European economies but also pose significant new risks, such as widespread cyber-attacks or large-scale accidents resulting from poor understanding of the systems' inner workings.

In response to these rapid developments, the EU Parliament recently approved the landmark AI Act in an effort to govern (advanced) AI systems. However, in its current form, the AI Act may be insufficient to curb the risks of models that will be released in the near future - even as soon as next year. Given the pace of improvement in AI capabilities, there is a critical need for more robust measures that leverage compute for effective AI governance. The following five recommendations suggest avenues to future-proof EU efforts.

The International Center for Future Generations (ICFG) is an independent think-and-do tank dedicated to shaping a future where decision-makers anticipate and responsibly govern the societal impacts of rapid technological change, ensuring that emerging technologies are harnessed to serve the best interests of humanity.

This brief is part of a broader series of State of Play reports on advanced AI scheduled for release throughout 2024. 'Advanced AI' refers to highly capable foundation models (such as GPT-4, Claude Opus and Gemini Ultra) or systems built on top of foundation models that can possess capabilities sufficient to pose serious risks to public safety. This particular document zeroes in on the role that computational power (or 'compute') plays in advanced AI developments. More specifically, it collects the most policy-relevant facts and trends on compute in AI and tries to present these in a non-technical manner[1]. Additionally, it provides five specific policy recommendations to assist EU policymakers in utilizing compute to foster responsible AI governance.

[1] The facts presented in this piece are based on an extensive literature review. Two sources have been particularly valuable for this writing: first, the work by Epoch AI on compute trends, and second, the recently published paper by Sastry et al. (2024) ('Computing power and the governance of artificial intelligence'). We thank Emily Gillett, Jaime Sevilla, Joshia Williams, Lennart Heim, Maxime Stauffer, Michael Aird, Tim Fist and Toni Laurence for their individual feedback on this brief. Readers are welcomed to reach out to advanced_ai@icfg.eu in case of questions or to discuss the policy recommendations made in this publication.

Key policy recommendations

01. Strategic allocation of EU compute resources.

The European High Performance Computing (EuroHPC) Joint Undertaking should shift its focus away from inadequate endeavors aimed at training competitive foundation models from scratch. Instead, EuroHPC's should double down on the other pillars of its AI Factories program by:

a. Enhancing the understanding, safety, and control of advanced AI models through compute-intensive research. This kind of research can also help spur a thriving European AI insurance sector.

b. Developing large but specialist AI systems that can help tackle societal problems in e.g. medicine, energy or climate science and which are not expected to be taken up by the leading advanced AI companies.

02. Extension of the AI Act's GPAI regulation.

The European Commission should prepare the addition of a third tier to the AI Act's GPAI regulation that addresses the severe systemic risks posed by the next generations of GPAI models. This extension would include:

a. An appropriate additional compute threshold above which GPAI models carry the presumption of severe systemic risk.

b. Requirements that mitigate the formation of dangerous capabilities during training and prevent pre-deployment proliferation of inherently hazardous model weights.

03. Compute-based enforcement scaling and prioritization.

The AI Office should scale and prioritize enforcement efforts in alignment with compute trends. More specifically, the AI Office should:

a. Strengthen its resolve to prioritize evaluation of GPAI models with the largest compute budget in case of (temporarily) limited personnel capacity.

b. Conduct or commission detailed capacity requirement projections based on compute trends, and hire/seek collaborations in line with those projections.

04. Establishment of an EU AI foresight unit.

The European Commission should create a dedicated AI foresight unit within the AI Office to better deliver on the Office's task to keep track of the evolution of AI markets and technologies. Studying (effective) compute trends would enable this unit to:

a. Discern several quantitative scenarios of future training compute budgets and inference capacities.

b. Work together with academia and civil society to map out what types of capabilities and accompanying risks might arise in the coming years for each of these scenarios.

05. Implementation of a multilateral compute oversight system.

The EU should start international dialogues to implement a multilateral compute oversight system that builds on the monitoring requirements of the EU AI Act and the recent US Executive Order 14110. This oversight system could start out as a bilateral agreement between the EU and the US and could afterwards be extended to other G7 countries. Monitoring requirements would:

a. Focus on the location of large AI clusters (theoretical maximum of >10^20 FLOP/s and >100 gbit/s networking) and planned or ongoing very large training runs (>10^26 FLOP).

b. Apply within each individual jurisdiction, with participating governments committing to sharing decision-relevant high-level information with each other.

Terminology

SYNTHETIC DATA:
Data that is not created by humans, often through the use of AI models.

FLOATING-POINT OPERATIONS (FLOP):
Roughly, the number of multiplications and additions performed on the AI chips.

ADVANCED AI:
Highly capable foundation models or (systems built on top of foundation models) that can possess capabilities sufficient to pose serious risks to public safety.

AI CLUSTER:
A collection of servers with AI chips that are connected through high-speed networks inside a data center.

RUNTIME COMPUTE:
The amount of computational operations used to serve an AI model to users.

CAPABILITIES:
Tasks an AI system can perform, like generating photorealistic images, or accurately answering scientific questions.

AI CLOUD SERVICE PROVIDER:
Third-party company that rents out AI clusters to AI companies.

EFFECTIVE COMPUTE:
The amount of compute required to train an AI model with certain capabilities in the absence of algorithmic improvements.

TRAINING:
The process by which an AI model is created using feedback from large amounts of data.

TRAINING COMPUTE:
The amount of computational operations used to train an AI model, often measured in floating-point operations (FLOP).

COMPUTE GOVERNANCE:
Governance of AI developments through the use of compute as a policy lever.

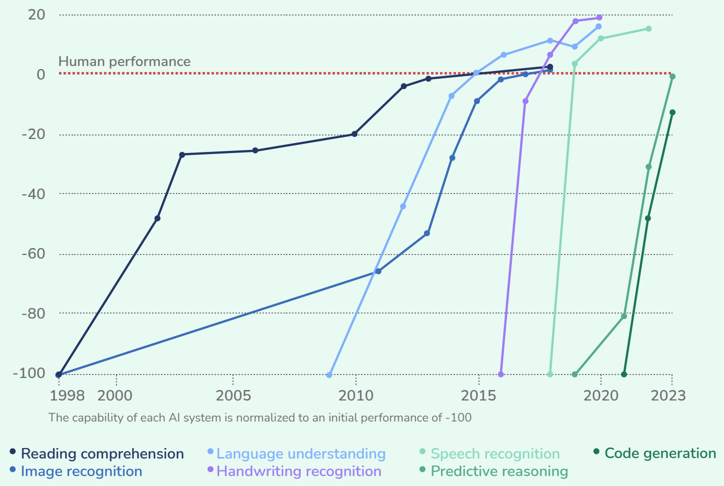
Introduction

Advanced AI systems like ChatGPT are rapidly reshaping industries, societies and lives. To steer this transition, the EU recently passed the AI Act - the world's first binding international legislation that establishes oversight and accountability for AI developers. Although a groundbreaking first step towards governing advanced AI systems, the AI Act in its current form may be insufficient to curb the risks of models that will be released in the near future.

The capabilities of AI systems are progressing at breakneck pace. In just a few years, language models have transformed from quirky research projects to productivity-enhancing tools that help millions of users brainstorm, draft reports, or write programming code. Simultaneously, the world was introduced to a host of new capabilities altogether, like text-to-speech indistinguishable from human vocals, or photorealistic image and video generation, sparked by a simple text-prompt. Similar capability jumps could well arise in the next couple of years.

The most important driver of this astounding progress in advanced AI has been the relentless increase of computing power - or compute - used to train AI systems. The largest training run in 2023 used approximately 10 billion times more computational operations than the largest training run in 2010 did - similar to the difference between a single human's effort and that of all humanity combined. With AI systems nearing capabilities that could render the technology truly transformative, it is essential that policy makers grasp the central role that this explosive compute growth plays in AI developments.

Compute is growing exponentially, and, alongside algorithmic innovation, will likely remain a key driver of progress in the remainder of this decade. Within the next few years, compute growth could already enable new AI capabilities such as highly skilled autonomous agents. These AI systems could revitalize European economies, but could also disrupt job markets, introduce new threats in cyber and biowarfare, or lead to large-scale accidents, for instance on financial markets[2].


AI has already surpassed human performance at a number of tasks

Source: Kiela et al. (2023) via Our World in Data

Figure 1: Advanced AI systems are acquiring new capabilities increasingly quickly.

[2] For more information about these risks, see the recent ICFG framing paper.

Autonomous AI Agents

› **AI companies are hard at work to increase the level of autonomy of their AI systems.** While current systems such as ChatGPT mostly respond to prompts, they may in the near future become able to sense their environment, autonomously make decisions and take more and more actions independently, especially in the digital domain. Think of ordering products on the web, sending emails, crafting complicated pieces of software consisting of multiple interlinking parts, or even entirely controlling your mouse and keyboard to take over digital workflows. AI systems capable of such tasks are often referred to as 'Agents'. There already exist AI Agents such as the AI software developer 'Devin', but these systems are still quite fragile and operate within relatively limited domains. Many experts believe that more powerful foundation models (e.g. GPT-5) will unlock better planning and more reliable execution of tasks, enabling AI Agents to fulfill their potential. This could spur the rise of very capable AI assistants that - like in the sci-fi movie Her - can help users with all sorts of increasingly complex tasks.

Sustaining historical training compute growth (the rate of which equals a staggering 4.2x per year) will become increasingly challenging for AI developers. To uphold current exponential growth rates until 2030 would likely require more than a 25x increase in yearly AI chip production compared to 2023. As advanced AI models grow larger, they also require increasingly more electrical power to run (both during training and during subsequent usage). Demand for electrical power from the global advanced AI industry is estimated to grow by an additional 75 GW by 2030 - on a yearly basis that is equivalent to roughly 25% of total power production in the EU. In response to these challenges, the AI industry is thinking big: current multi-billion dollar plans by AI companies and their partnering cloud providers lay out avenues to uphold compute scaling for another 5 years.

Because compute is such an important determining factor in AI progress, it also provides an essential pathway to effective AI governance. Regardless of whether its growth slows down, compute has unique properties that make it a very attractive policy lever: compute cannot be copied, it is relatively easily detectable and quantifiable by outside actors, and has a very concentrated supply chain. The same cannot be said for data and algorithms, the other two essential AI training ingredients (while there may be limiting factors on those, they do not provide the same governance opportunities to democratic institutions). Given these properties, the EU should urgently invest capacity in compute governance. By using compute as a lever for governance, the EU can improve public oversight over AI developments and can strengthen enforcement of key AI regulations such as the EU AI Act.

This report provides a technical briefing of compute's role in AI development. It subsequently presents five concrete policy recommendations for how the EU can harness compute as a policy lever to ensure that AI serves the public interest, and becomes a strengthening force for democracy, the rule of law and human rights.

Technical briefing

Compute is an essential ingredient for training advanced AI systems

AI systems are trained through feedback from data, not programmed directly.

AI models can be considered software, just like the software running on your phone. They are, however, entirely different kinds of software. Classic software is meticulously written by human programmers and put together like a recipe: if situation X occurs, do Y; if not, do Z. By contrast, AI models are trained using an automatic process in which a model receives feedback from data. It is almost like the model is grown. First, human programmers specify an algorithm that is capable of learning but which does not yet hold any knowledge about the world. From there, the model is shaped organically by exposure to data, forming concepts and latent skills along the way without human interplay. The advantage of this approach is that the model's learning curve is not limited by human speed or ingenuity. The disadvantage is that even developers do not really understand how advanced AI systems work internally: they are essentially black-box systems whose behavior is difficult to predict. Current 'interpretability' techniques provide some insights into the internals of advanced AI systems, but are years away from the current frontier. For example, with the help of GPT-4, OpenAI was able to locate some of the internal 'concepts' formed in GPT-2, but with low accuracy. GPT-2 is 10,000 times smaller than current state-of-the-art models.

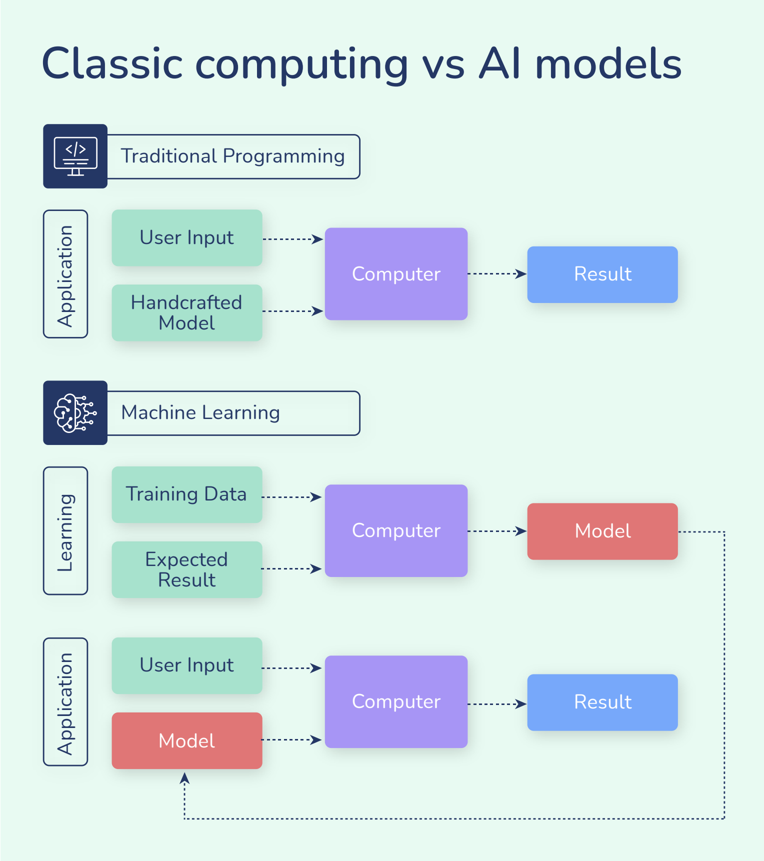


Figure 2: AI systems are trained through feedback from data, not programmed directly.

The AI triad: data, algorithms and compute.

Training an advanced AI model requires three ingredients: data, algorithms and compute. Data is the material that provides the model with feedback during training. Algorithms are used to process the feedback the model receives from the data. Compute is what ties the two together: specialized chips are typically required to run the algorithms that process the data. Human capital serves as an underlying driver for these three inputs. Figure 3 summarizes this 'AI triad'.

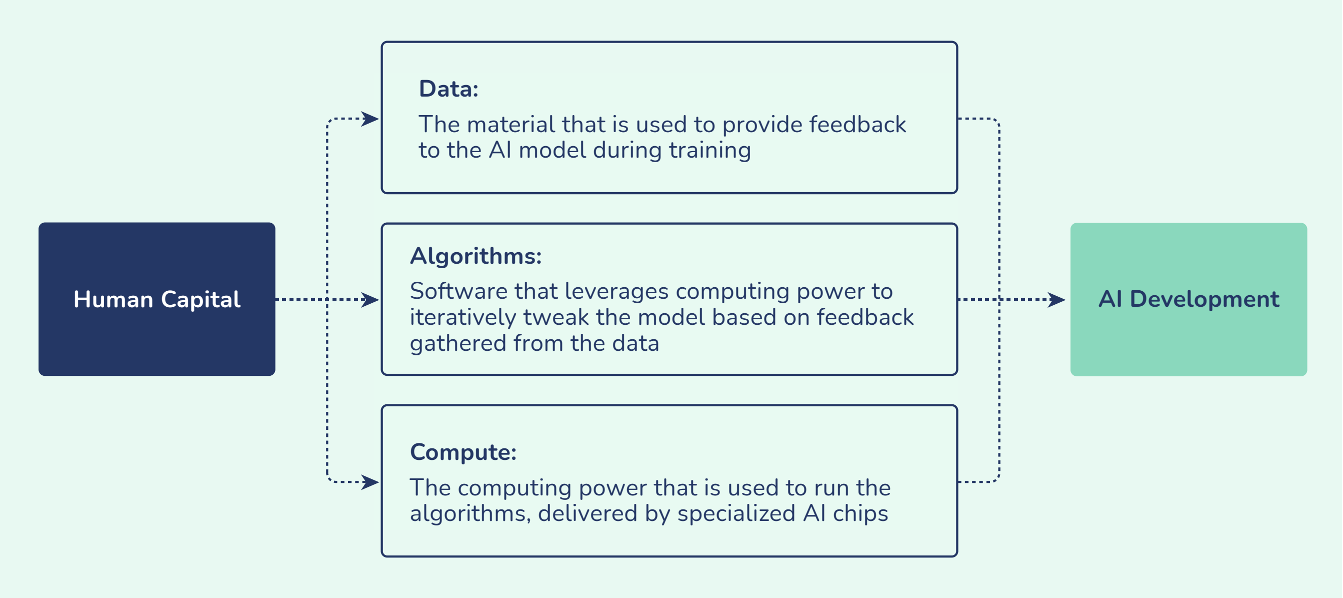

Human Capital
Data:
The material that is used to provide feedback to the AI model during training
Algorithms:
Software that leverages computing power to iteratively tweak the model based on feedback gathered from the data
Compute:
The computing power that is used to run the algorithms, delivered by specialized AI chips
AI Development

Figure 3: Training an AI model requires three ingredients: data, algorithms and compute

Compute scales with data quantity and model complexity

Compute is most commonly defined by the number of so-called floating-point operations (FLOP) used to train a model[3]. Floating-point operations roughly measure the number of multiplications and additions performed on a chip. So, for example, to multiply two pairs of numbers and to subsequently add up the result, would require a total of 3 FLOP. Training advanced AI models requires vast numbers of such multiplications and additions. These training workloads are enabled by large numbers of specialized AI chips, each capable of performing trillions of FLOP every second[4]. If AI companies acquire access to more or faster chips, they can perform more floating-point operations. These additional floating-point operations can either be used to make the model more complex, or to train the model on more data points. Typically, a combination of the two is most efficient, which is why AI developers are collecting and generating increasingly more data to train their models. For instance, the recently released DBRX model by Databricks was trained on almost 10 trillion words - a significant fraction of the text available on the internet.

Serving AI models to users also requires compute at runtime

After an AI model is trained, it can be deployed to serve users. Serving an AI model at runtime also requires compute, but much less so than during training. To see why, note that during training, a language model has to process billions to trillions of words[5], whereas a request from a single user often only requires a couple of hundred words to be processed. However, as the number of users and their number of daily AI interactions grows, so does total compute used at runtime. It is therefore possible that runtime compute will eventually outgrow training compute. In anticipation of user growth, AI companies are already pursuing ways to decrease inference compute requirements by training their models for longer.

[3] The appropriate unit depends on the training method. For instance, compute can also be expressed in the number of integer operations used during training.

[4] For instance, the NVIDIA H100 AI chip is able to perform 1979 trillion FLOP per second, in PF8-precision.

[5] Or images, videos, audio, etc.

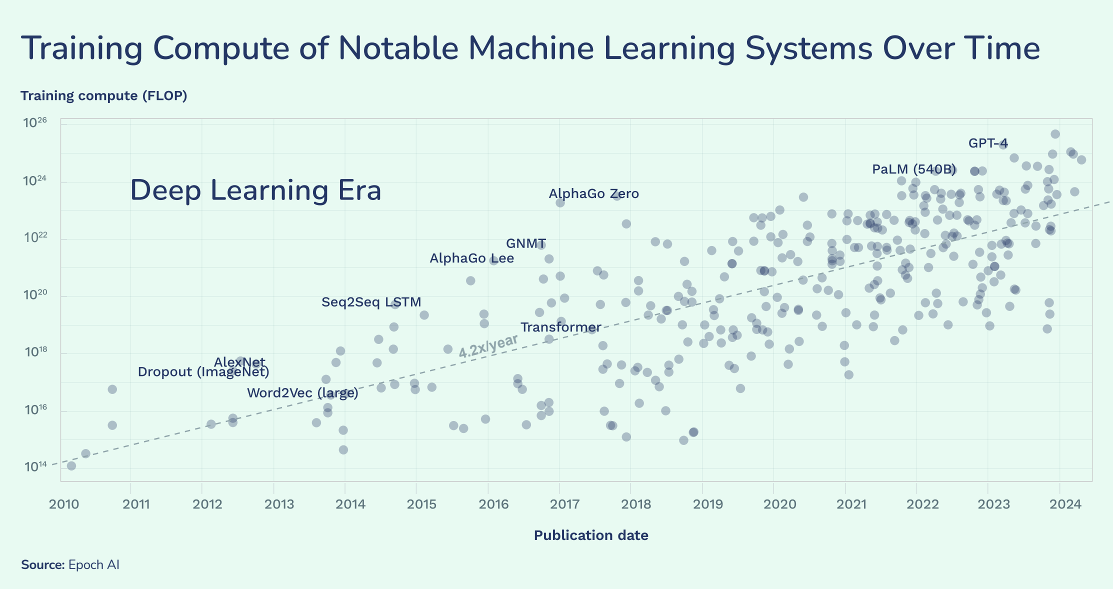
The exponential increase of training compute has been the predominant driver of recent progress in AI

Compute has been growing exponentially at four times the growth rate of Moore's law

In recent years, the number of floating-point operations used to train AI systems has skyrocketed. Some 10^15 FLOP were performed to train the leading AI model in 2010, whereas 13 years later, Google's Gemini Ultra was likely trained for more than 10^25 FLOP. That's a million times more than the estimated number of sand grains on earth.

Data from Epoch AI shows that training compute has been increasing by a staggering 4.2x per year since 2010. This is much faster than Moore's law: each time the number of transistors on a chip doubles, the compute used to train advanced AI systems increases by a factor of 18x. Figure 4 presents this trend (note the logarithmic y-axis).

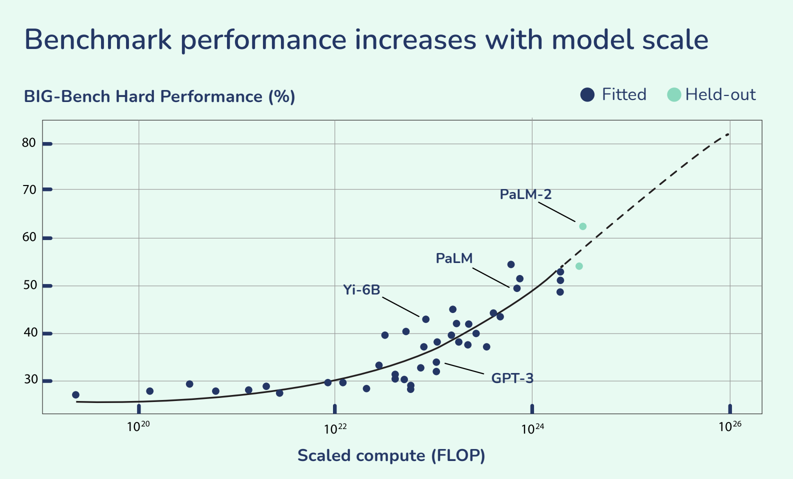
Figure 4: Compute used for training large AI-systems since 2010, logarithmic scale.



Increases in compute stem from both improvements in computational price performance (faster chips for the same price) and increases in spending. Spending growth has been the main motor behind the exponential growth in training compute. Epoch AI estimates that computational price performance increases by a factor of 1.4x every year, whereas spending on training runs increases by a factor of 3.1x per year. The large contribution of spending increases explains how compute has been able to grow faster than Moore's Law.

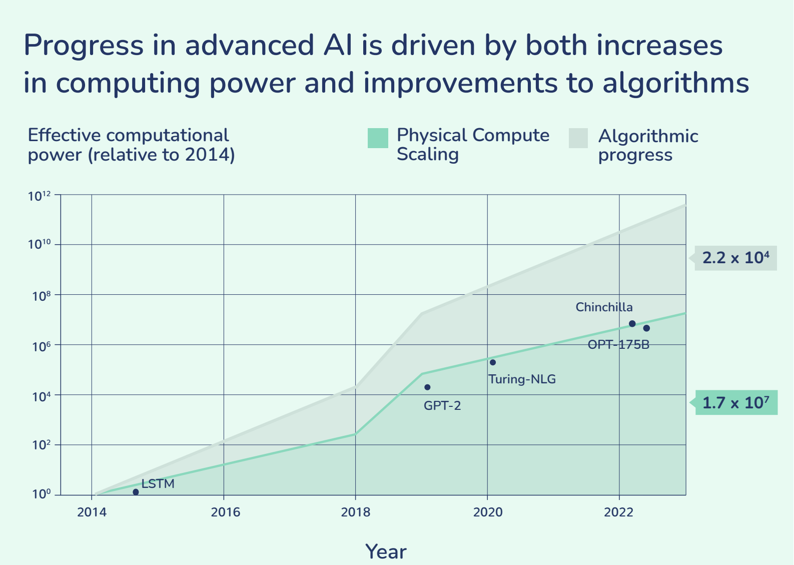
'Scaling laws' describe how AI systems improve when trained with more compute

The exponential growth in compute is reinforced by the recent discovery of so-called scaling laws. These are empirically observed relations which reveal that performance on training tasks, such as the AI system's next-word-prediction accuracy, improve systematically when developers spend more compute to train a model. However, the exact relationship between compute and downstream



Source: David Owen. 'How predictable is language model benchmark performance?'. ArXiv [cs.LG], 2024. arXiv. https://arxiv.org/abs/2401.04757.

Figure 5: Downstream question-answering capabilities as measured by the BIG-Bench benchmark scale with compute but are hard to accurately predict.



Source: Epoch AI

Figure 6: Algorithmic progress has contributed significantly to recent progress in advanced AI.

capabilities, such as question answering, is more poorly understood. Developers can expect the model to become more capable by increasing training compute, but cannot reliably predict in what ways the model will become more capable (e.g. it cannot be predicted in advance if the model will suddenly learn economically useful or dangerous new skills, or what skills will be most improved). OpenAI CEO Sam Altman even went so far as stating that predicting downstream improvements is a 'fun guessing game'.

Progress in AI is compounded by algorithmic improvements

Progress in advanced AI is not only driven by compute, but also by improvements to algorithms[6]. Research by Epoch AI shows that, on average, AI systems in language modeling and vision need approximately 3x less compute to reach the same quality as last year's systems, as a result of algorithmic progress. Between 2014 and 2023, algorithmic progress in pre-training has enabled AI model performance to improve as much as it would have with around 22,000x more compute. Together, compute and algorithmic improvements account for more than a 10x yearly increase in developers' effective resources. The compounding effect of this number can hardly be overstated: between 2014 and 2024 so-called 'effective compute' increased by more than a factor 100 billion.

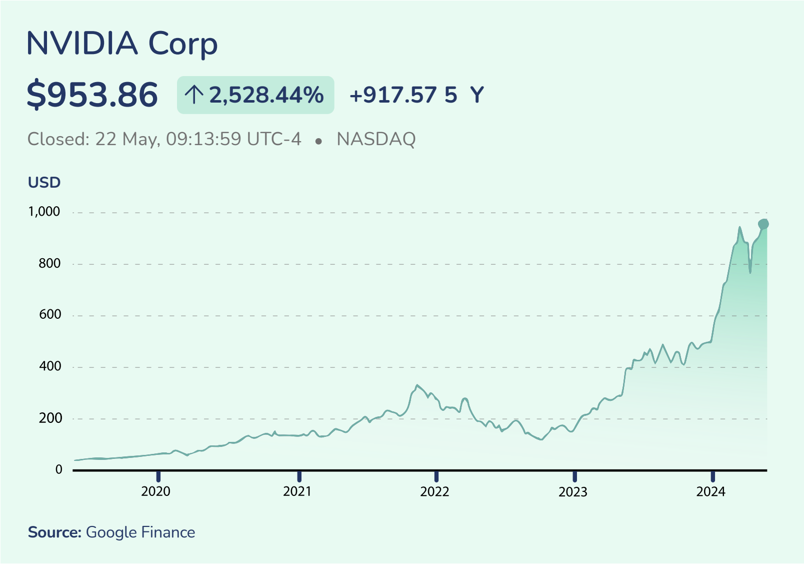
Developers are betting they can keep turning compute into new AI capabilities

If progress in AI continues at its current pace, highly capable autonomous AI agents may be developed within the next few years[7]. Such agents could upend European economies by automating labor-intensive tasks and by speeding up science and R&D. However, the same systems could also disrupt the job market, or be misused to commit large-scale terrorist attacks if governed irresponsibly. Even if specific

[6] Here the term 'algorithmic progress' should be interpreted quite broadly, to also encompass improvements in data quality and hardware utilization.
[7] Such agents are already pursued widely in the AI industry with increasing success. See for instance, the recently released AI agent 'Devin' that can autonomously complete complex software engineering tasks.

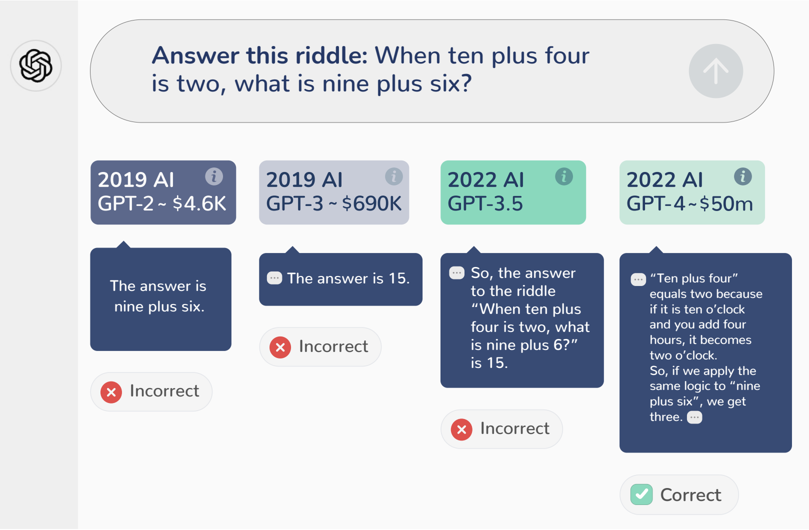
predictions of AI impacts turn out wrong, the sheer size of investments in generative AI highlight that it is a sector on the rise. Bloomberg estimates that generative AI alone can become a trillion dollar market by 2032. AI companies are racing to create a range of new capabilities that could prove transformative - and they are betting on compute growth to unlock these advances.

There is concrete evidence that leading AI companies are continuing to ramp up their compute investments in line with historic growth rates. For instance, Mark Zuckerberg recently stated that by end of 2024, Meta will have access to the equivalent of 600,000 NVIDIA H100 chips (a current state-of-the-art AI chip). If correct, Meta's total compute base will be almost 100 times larger than the cluster GPT-4 was trained on, although these chips will be spread over different clusters. Meta's accompanying hardware investment will likely exceed 10 billion USD[8]. Furthermore, Microsoft is reportedly building a single multi-site training cluster for OpenAI that consists of 'hundreds of thousand' AI chips, and Amazon has just bought a datacenter location with a dedicated 1 GW nuclear power plant that could provide electricity for a cluster of almost 1 million AI chips. OpenAI and Microsoft are even reported to have begun planning a build-out of a 5 GW AI supercomputer called 'Stargate' that would host 'millions of AI chips'. This massive cluster is supposedly planned to be operational between 2028 and 2030 and could likely be used to train models on the order of 10^29-10^30 FLOP - models 10,000x to 100,000x more compute-intensive than GPT-4. Perhaps unsurprisingly, NVIDIA's stock has skyrocketed this year, largely due to immense AI chip orders from big tech companies like Microsoft, Google, Amazon and Meta (see Figure 7).



Source: Google Finance

Figure 7: NVIDIA, the leading AI chip design company, saw its stock skyrocket after the launch of ChatGPT - stock price on March 29th, 2024.



Answer this riddle: When ten plus four is two, what is nine plus six?

2019 AI GPT-2 ~ $4.6K: The answer is nine plus six. ☒ Incorrect

2019 AI GPT-3 ~ $690K: The answer is 15. ☒ Incorrect

2022 AI GPT-3.5: So, the answer to the riddle "When ten plus four is two, what is nine plus 6?" is 15. ☒ Incorrect

2022 AI GPT-4~$50m: "Ten plus four" equals two because if it is ten o'clock and you add four hours, it becomes two o'clock. So, if we apply the same logic to "nine plus six", we get three. ☑ Correct

Figure 8: OpenAI's GPT model family has steadily improved over time with growing compute expenditures.

There have been no empirical results suggesting that scaling laws will soon peter out, let alone hit a brick wall[9]. In a recent interview, Dario Amodei, CEO of Anthropic stated that he personally thinks 'it is very unlikely

[8] Assuming an average price of 20,000 USD per H100-equivalent yields a total of 12 billion USD. This may even be an underestimate as H100 prices have reportedly been as high as 30,000 USD per chip.

[9] Note that academic research on the limits of compute scaling is mostly missing due to poor access to compute relative to the corporate AI industry.

that the scaling laws will just stop'. Amodei has been a long-time advocate of the so-called 'scaling hypothesis' and correctly predicted that more compute would increase capabilities as early as when OpenAI developed their first GPT-model. Historical trends indicate that training compute grows by a factor 4.2 every year - this is roughly the same compute difference as the jump from OpenAI's ChatGPT-3.5 to ChatGPT-4. While GPT-3.5 scored in the bottom 10% of a simulated bar exam, GPT-4 scored in the top 10%. If compute and algorithmic progress continues to scale in this tempo, we could expect similar jumps in capabilities every year.

Even if scaling laws do peter out, compute would still remain an important input to AI progress. After all, compute is not only used to process data during the training stage: it is also essential for algorithmic experimentation, to curate or enhance data sets, and to let models 'think for longer' before they answer a user's request, so as to improve output quality (see Text Box 1).

Compute can drive AI progress even if scaling laws peter out

Compute is essential for algorithmic progress. Unlike many other scientific disciplines, breakthroughs in advanced AI mostly do not result from theoretical insight, but rather, from simply trying things out. The more compute a developer has access to, the more algorithmic experiments they can run in parallel, speeding up progress. As advanced AI systems become more complex and data-hungry, developers require increasingly large test runs to ascertain whether a proposed algorithmic adjustment really improves model performance[10]. Indeed, OpenAI's former head of developer's relations, Logan Kilpatrick, recently stated that OpenAI's core research team is deliberately kept small, since algorithmic progress is constrained by compute availability, and not by the number of employees pursuing innovative algorithmic adjustments. In a recent interview, Demis Hassabis, CEO of Google Deepmind, said: 'We use our compute not just for scaling. You need quite a lot compute to do new invention because you've got to test many things at least at some reasonable scale. Some new ideas may not work at toy scale but may work at larger scale. And in fact those are the more valuable ones.'

Besides experimentation, compute can also be leveraged to generate so-called synthetic training data: text generated by an AI model that serves as the training data for another model[11]. Recent research by Microsoft shows that training models on carefully curated and synthetically enhanced datasets can yield substantial efficiency benefits: using synthetic data, a more capable model can be trained with the same amount of compute. Leading AI companies are currently investing heavily in synthetic data approaches - not only to improve data quality, but also to increase data quantity, as developers are quickly running out of existing internet data to train their models on.

The last two years have also seen an explosion of new methods that use clever ways of turning so-called 'runtime compute' into better model performance. Slightly simplified, these methods enable the model to 'reflect' and 'think for longer' before answering a request, which requires more compute, but also yields higher quality output. In the near future, runtime compute methods may also be used to generate very high quality synthetic data that next-generation models can be trained on to intuitively grasp. Using the enhanced 'intuition', the next-generation model can produce even higher quality output when prompted to think for longer and reflect on its output. This so-called bootstrapping approach may enable developers to iteratively improve their models, but would also require enormous amounts of compute.

[10] For instance, Meta recently released a new paper showing that so-called multi-token prediction can yield significant efficiency gains but that these benefits really only kicked in for models larger than 13 billion parameters.

[11] Note is also possible to create other types of synthetic data, like images or audio.

But it will be increasingly difficult for AI companies to uphold growth rates in compute

So far, the exponential growth of training compute has been very reliable. However, as experts have pointed out, it will be increasingly difficult for AI companies to uphold the current exponential. First of all, AI companies can bump into spending constraints: even Google, Microsoft, Amazon and Meta have finite budgets, and it could become hard to justify exponential CAPEX increases to shareholders, especially if the next generation of AI systems does not constitute a big leap forward.

Second, AI companies might run out of data to efficiently keep scaling their models. Epoch AI estimates that developers could already exhaust high-quality text-data during 2026. That said, AI companies are hard at work to overcome this data bottleneck by making use of other types of data (images, videos, audio) and by training on synthetic data generated by other AI systems[12].

AI companies would also have to circumvent multiple physical bottlenecks. For instance, to uphold current exponential growth rates until 2030 could require more than a 25x increase in yearly AI chip production compared to 2023[13]. At that point, AI chips would take up half the production capacity as today's high-end chips for mobile, PC and cloud combined. Note that supply constraints are no mere hypothetical: in their recent earnings call, NVIDIA stated that they expect to be supply-constrained throughout 2024, even though chip fabrication firms such as TSMC are rapidly expanding their highly specialized production lines.

AI's future power demand could put great stress on the world's ability to generate enough renewable electricity to keep climate change in check

Another potential future bottleneck is electrical power. The next-generation NVIDIA NVL-72 server - which contains 72 specialized AI chips plus supporting hardware - will draw 120 KW of power and must be liquid-cooled to prevent overheating. That's almost 2000 W per AI chip, and AI companies such as OpenAI are reportedly building (multi-site) clusters of hundreds of thousands of AI chips for training purposes and to serve their rapidly expanding user base. As a consequence of this immense scaling, demand for electrical power from the global advanced AI industry is estimated to grow by an additional 75 GW by 2030 - equivalent to 25% of total power generation in the EU on a yearly basis. This power demand would be concentrated in specific regions (mostly in the US where energy does not have to be imported and where electricity prices are relatively low), potentially leading to local or regional supply bottlenecks.

AI's future power demand could put great stress on the world's ability to generate enough renewable electricity to keep climate change in check, and could necessitate that advanced AI systems themselves contribute to breakthroughs in energy production, storage or negative emissions technologies. Besides power, cooling all these chips requires serious amounts of water. Although global water supply is plenty large enough to fulfill this demand, local water shortages could be exacerbated by exponential growth of the data center sector.

[12] For instance, Google's Gemini Ultra 1,0 was trained on text, images, audio and video.

[13] 4,2^6 equals roughly 5000. However, AI chips could get roughly 10 times faster during the next 6 years and training run length could probably increase by another factor 5. Furthermore, it seems likely that a larger proportion of total chip production will go to the largest training run in 2030, as a result of winner-takes-all dynamics. Assuming a fourfold larger concentration yields a total production factor of 25.

AI industry leaders seem to be recognizing the magnitude of these challenges. Microsoft and Amazon are looking at nuclear plants to power their AI data centers and NVIDIA has reportedly closed a deal with Intel Foundries to package an additional 3 million AI chips per year (TSMC is currently packaging nearly all AI chips and cannot keep up with demand despite heavy investments in new production facilities). Meanwhile, OpenAI's CEO Sam Altman is supposedly trying to spur enormous investments in the power- and semiconductor industries, with quoted numbers up to seven trillion USD - higher than Germany's entire annual GDP.

Regardless of whether OpenAI can pull off such an immense challenge, the buildout of the aforementioned Stargate supercomputer alone would probably be sufficient to uphold compute trends until 2029. In the near future, it therefore seems likely that compute will remain a key driver of progress in advanced AI. Furthermore, five years of continued exponential growth may be long enough to enable truly transformative AI systems. Although it seems virtually certain that developers will eventually run into bottlenecks, policy makers should not conclude that compute-driven progress will be negligible.

Policy Recommendations

Why compute is a useful lever for AI governance

Compute is excludable, detectable, quantifiable and has a concentrated supply chain

Compute has been the predominant driver of recent progress in AI, and will likely continue to be an important driver of progress in the coming five years. But there's another reason policy makers should pay attention to compute: compute is a very promising lever for AI governance. It is excludable, detectable, quantifiable and has a very concentrated supply chain - all properties conducive to effective governance:

- **Excludability:** AI hardware cannot freely be copied, unlike algorithms and data.
- **Detectability:** AI companies need large amounts of power and space for big data centers, whereas training data and algorithms can be stored on a single hard drive.
- **Quantifiability:** Compute can be relatively easily measured, reported and verified, whereas data quality and algorithmic efficiency are harder to quantify.
- **Concentrated supply chain:** There are only a handful of players who design, produce and provide access to AI chips, whereas there are many active players who produce data sets and algorithms (see Figure 9).

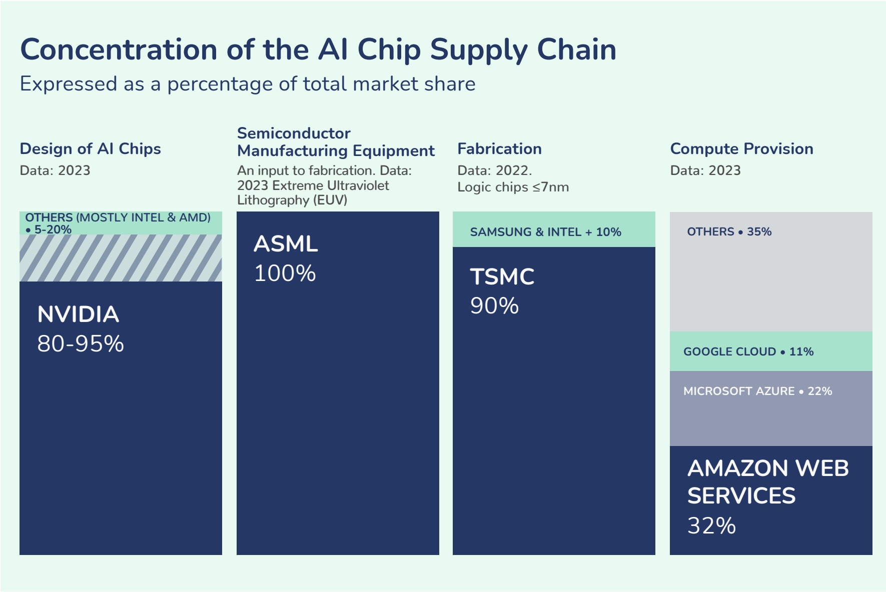


Figure 9: The AI supply chain is very concentrated, with individual companies often owning the majority of a segment's market share.

Due to these properties of compute, we are already seeing the first signs of so-called compute governance in the wild. The EU AI Act, for instance, classifies general purpose AI systems by the compute that went into training them: systems trained on more than 10^25 FLOP, in a compute league of their own, are presumed to belong in the systemic risk category. However, there are more ways the EU can harness compute to ensure that AI serves the public interest, and strengthens Europe's open markets and societies.

How knowledge and regulation of compute can improve AI governance: five recommendations

The EU can use compute as a lever to further improve AI governance in five distinct ways:

1. By investing in compute in a targeted way to promote our understanding and control of advanced AI models, and to train specialized AI systems that can help tackle large societal issues in e.g. medicine, energy and climate science.
2. By preparing the addition of a third tier of regulation for GPAI models with severe systemic risk, using a classification mechanism that is partially based on training compute.
3. By strengthening the AI Office's resolve to prioritize evaluation of the most compute-intensive GPAI models in case of (temporarily) limited personnel capacity and to increase capacity in line with compute trends.
4. By adding a dedicated foresight unit to the AI Office that helps anticipate future AI policy challenges based on (effective) compute trends.
5. By leading the way on establishing a multilateral compute oversight system that can aid each of the previous 4 efforts.

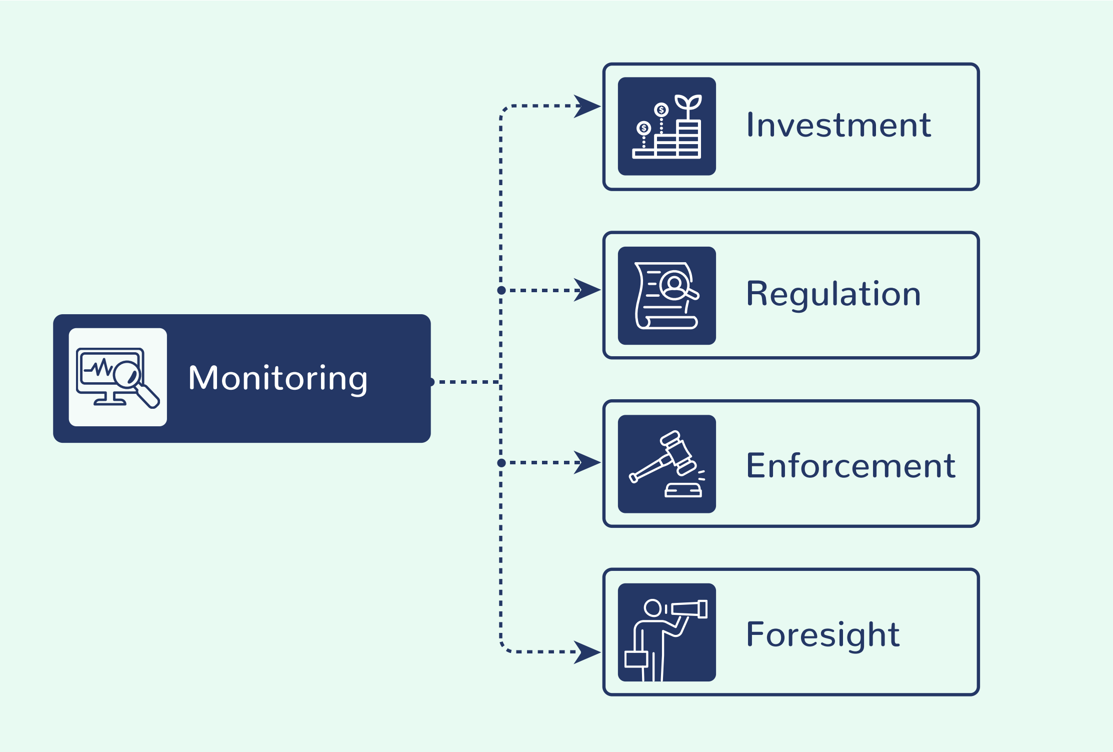


Figure 7: Compute monitoring can aid compute investment, regulation, enforcement and foresight.

RECOMMENDATION 1

Redirect compute investments to AI safety research and training of large, specialized models to solve important scientific problems

Through the EuroHPC AI Factories, the EU will provide AI compute to European initiatives that aim to 'train large General Purpose AI (GPAI) models'. However, given the vastly superior compute resources of private US AI companies, it seems increasingly unlikely that these investments will suffice. Instead, the EuroHPC Joint Undertaking should double down on the other pillars of the AI Factories program, such as 'testing, evaluation and validation' of large scale GPAI models, and the development of AI-solutions for 'science problems'. By deliberately allocating compute resources to compute-intensive AI safety research the EU can have an outsized influence over global AI developments and lay the groundwork for a thriving European AI assurance industry, all the while incurring limited costs. Similarly, by supporting initiatives to train large but specialized models to be used in for instance vaccine development, material science or self-driving, the EU can spark scientific breakthroughs that can invigorate the EU economy.

Currently, the field of advanced AI is dominated by a handful of private players, mostly based in the US. The incentives these companies face need not always align with EU values and the winner-takes-all dynamics in advanced AI could bring about severe concentration of power. Academia is no longer on the forefront of advanced AI, largely due to lack of compute access (the discrepancy between the compute available to private companies and academia is often called the 'compute divide'). Given this state of affairs, it is commonly suggested that the EU steer the development of AI by investing public funds in its own corporate European AI champions.

However, as the recent partnership between Microsoft and the French AI start-up Mistral AI shows, it is very difficult to compete in advanced AI without backing from American cloud service providers. Preventing the accompanied concentration of power would require heavy public investments, and reallocating existing compute resources simply won't cut it. As an example, the European Commission has recently opened access to a selection of European supercomputers for AI startups. Combined, the eight EuroHPC supercomputers house some 32,000 specialized AI chips, most of which are previous generation NVIDIA chips. Microsoft is reportedly targeting 1.8 million AI chips by the end of 2024, with a much larger percentage of those being state-of-the-art chips. That's roughly a 100x difference in AI compute resources.

Recent amendments to the EuroHPC programme are a step in the right direction, but still too small and scattered if the goal is to help European AI start-ups

By the end of 2024, Microsoft will have access to roughly 100x more AI compute than all eight European supercomputers combined

compete directly with the leading US AI companies. Through the EuroHPC Joint Undertaking, the EU will redeploy an existing 2.1 billion USD of funding towards AI Factories, with a specific aim to enable large GPAI model training. However, given the distributed nature of the EuroHPC program, this investment will be allocated over multiple data centers and countries. Under the optimistic assumption that a total of 500 million euros will be spent on a single AI compute cluster, this still only buys some 20,000 NVIDIA H100's. Once the AI Factories are fully operational, OpenAI, Google and Meta will probably have access to clusters that are 10-100x larger. Moreover, the leading US AI companies have already attracted most of the world's leading AI talent, have years of proprietary algorithmic innovation at their disposal, and have already collected heaps of user-data to further refine their models on. European competitors are thus heavily behind on multiple fronts, and catching up on foundation model training seems near-impossible without much larger funding or cooperation with the American cloud-service-providers. Furthermore, as long as companies like Meta are open-sourcing competitive foundation models (like the recently released Llama-3), European companies can just build on these public goods, instead of reinventing the wheel.

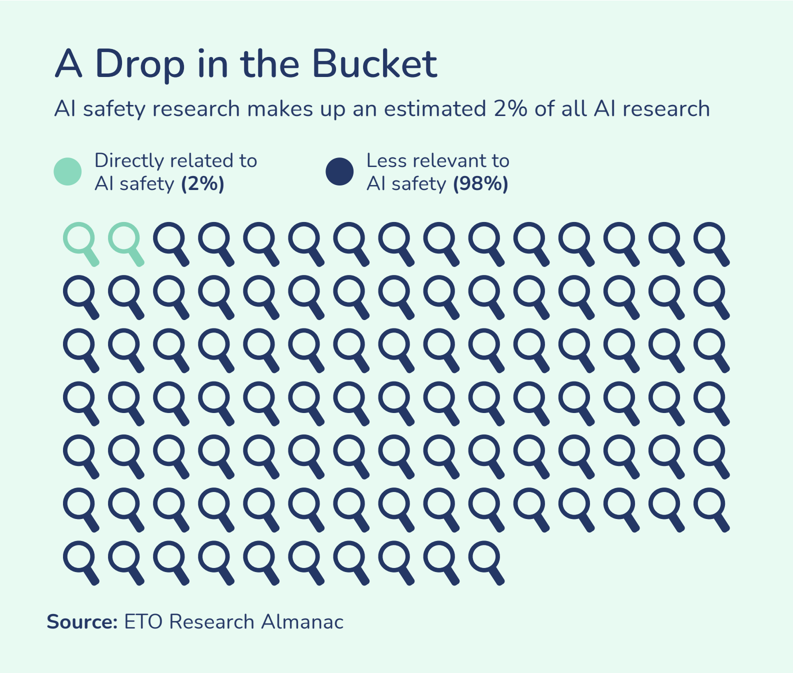
If the window of opportunity to compete directly on advanced AI has already closed, the EU should consider pursuing a different investment strategy to remain a relevant player. One promising option is to fund compute-intensive efforts that:

1. Improve our understanding, safety and control of AI systems through experimentation with large open-source foundation models or via structured access to closed models.

2. Aim to build large but specialized models that can help tackle important societal problems like vaccine development, material science for efficient batteries, or for modeling local effects of climate change.

As corporate developers are getting more entangled in a race to outcompete each other, they will face increasingly large incentives to cut corners on safety. Moreover, the leading AI companies have not yet made much progress on so-called 'scalable alignment techniques': ways of ensuring a model behaves as intended that can be scaled to arbitrarily large compute budgets. Public funds may thus be required to move the needle from 'fun guessing games' to predictable and safe AI designs. AI safety research is heavily underfunded and while it often requires significant compute resources (see e.g. Anthropic's work on dictionary learning), compute requirements are smaller than for training competitive foundation models from scratch. The EuroHPC AI Factories are thus ideally suited for this type of work. Pooled funding and structured access to state-of-the-art foundation models could further enable automated, large-scale testing of model's guardrails against misuse, and help create reliable evaluation tools. Only some 2% of AI research is currently focused on safety research (exact numbers depend on the methodology chosen).

[14] Lumi houses 11.912 MI250x chips, LEONARDO 13.824 A100 chips, MareNostrum5 4.480 H100 chips, Meluxina 800 A100 chips, Karolina 572 A100 chips and Deucalius 132 A100 chips.


A Drop in the Bucket

AI safety research makes up an estimated 2% of all AI research

Source: ETO Research Almanac

Figure 8: AI Safety research comprises just a fraction of total AI research.

Reallocation of the EuroHPC resources towards research on topics such as AI interpretability, reliability and evaluation would likely be sufficient to double the number of projects existing in those areas today[15].

Such research could not only improve the safety of models that are created overseas - it could also bring about immense economic benefits for Europe. The poor reliability of AI systems is one of the prime reasons for their current lack of adoption, and if AI is to make up on its economic promises, the world needs to prevent accidents. Public investments could also spur a thriving ecosystem of European AI assurance companies, the market for which is just beginning to grow.

To further strengthen the European AI ecosystem, the EU can invest in applied AI research that requires large, but specialized models. Specialized AI models have already been used to speed up vaccine development and discover new materials that could one day be used in ultra-efficient batteries. However, the leading AI companies have only pursued these types of specialized models sporadically (Google Deepmind is a notable exception here with groundbreaking results such as AlphaFold and GraphCast). The limited attention from the leading firms opens up opportunities for European researchers and corporate spin-offs to tackle science problems that unlock immense economic and social benefits. These problems typically require large compute resources, but not as large as training general-purpose models. It is thus a less capital-intensive market where competition is much less fierce, and where European players can more easily secure a competitive advantage. In the space of large, specialized models there is also ample room to use EU resources to stimulate the responsible development and deployment of open-source models that the European research- and startup communities can build upon.

[15] Rough estimates indicate that currently only some 300 people are working on technical AI safety.

RECOMMENDATION 2

Add a third GPAI tier to the AI Act for models with severe systemic risk

To future-proof the AI Act, the Commission should prepare the addition of a third tier of regulation for very large GPAI models that adds pre-development safeguards to the existing set of safeguards for GPAI models with systemic risk. Currently, the EU Act's requirements for GPAI models with systemic risk rely on corrective measures: if an AI company releases a model that does not comply with Article 55, the AI Office can demand the company take mitigating measures, or in the worst case scenario, issue the removal of the model from the EU market. Such post-deployment correction will likely suffice for current-generation GPAI models with systemic risk. After all, the amount of harm that can be done during the pre-correction time frame is fairly limited given today's model capabilities. Even if a non-complying model would proliferate, the negative impacts seem bearable - there are already uncensored versions of reasonably powerful GPAI models out there on the internet that so far have not caused mayhem. There is also reason to believe that non-compliance will be limited: AI companies will generally not want to lose out on the EU market, so the AI Act creates strong incentives to comply with regulation. However, the AI Act's corrective approach could fail to keep EU citizens safe from the next generations of advanced AI models. This is because corrective measures do not fully protect against accidents, and the severity of possible accidents is bound to increase.

The AI Act's current corrective approach could fail to keep EU citizens safe from the next generations of advanced AI models

To make this case more concrete, note that under current legislation, an AI company could theoretically train a 10^26 FLOP class model without the AI Office knowing (e.g. a model 10x more compute-intensive than GPT-4). After all, the AI company may train the model for the 'sole purpose of research, development and prototyping activities' and will therefore not face any reporting requirements. If the AI company and its cluster provider fail to take precautionary infosecurity measures, the model could be stolen by hackers, criminal organizations or adversary states in the final stages of training. If this seems far-fetched, note that insiders at leading US AI companies are quoted saying their infosecurity is so bad that 'they are doing more to accelerate the AI capabilities of US adversaries, than the adversaries themselves are'.

If the model's parameters are uploaded to a torrent website - as happened after the unintentional leak of a recent Mistral model - it would be virtually impossible to take the model offline again (there is no way to get rid of all local copies simultaneously) and prevent its potential guardrails from being fine-tuned away (this happened within a day after Meta's Llama-1 leaked). The same model could then be used by anyone to spread disinformation at unprecedented scales or potentially perform automated cyber-attacks on EU targets. The EU would thus face irreversible proliferation of very capable and dangerous GPAI models. To make things worse, at no point in this chain of events would the AI Office necessarily learn which company trained the AI model, or who owned the compute cluster it was trained on. It may therefore be impossible to hold the parties involved liable. As such, neither the AI company nor the cloud service provider faces proper incentives to strengthen their infosecurity.

Relying on addressing issues only after training has finished, becomes less and less viable as GPAI models grow more powerful

Alternatively, a GPAI developer may accidentally train a next-generation model that learns deceptive behavior and develops the necessary situational awareness to hide this behavior during training and evaluation. Note that humans do this all the time: consider someone driving very carefully and responsibly during lessons and examinations, only to start speeding after they acquire their driver's license. In fact, there are studies showing that today's frontier models are also capable of 'playing the training game' and that it's really hard to spot and remove these dangerous tendencies when they arise. With companies like Meta, Google and Microsoft embedding their latest models into products that are used by billions of users, the deployment of such deceptive models could cause large-scale harm in a matter of days, and might in the future even lead to loss-of-control scenarios.

We can draw a general lesson from these examples: relying on addressing issues only after training has finished, becomes less and less viable as GPAI models grow more powerful. The EU should thus add a third, enforceable tier of regulation for GPAI models that adds pre-development guardrails for GPAI models with severe systemic risks. Classification mechanisms could build on the existing ones: as long as proper pre-development risk-assessments are lacking, GPAI models could come with the presumption of severe systemic risk if trained using more than 10^26 FLOP. As in the current AI Act, the Commission should further be able to classify a model below this threshold as having severe systemic risk if the model has capabilities equivalent to the first generation of 10^26 FLOP models (due to algorithmic progress, future models could reach similar risk-levels using less compute). In practice, such a threshold would uniquely target models trained on giant compute clusters (roughly equivalent to 30,000 NVIDIA H100 GPUs or more, capital expenditures for which exceed 500 million USD), so would only apply to a handful of AI companies. A 10^26 FLOP threshold would target models in the future GPT-4.5 to GPT-5 class - i.e. models 1 or 2 generations down the line. Given that GPT-4 class models are already capable of lowering the threshold for the creation of bioweapons (although not so meaningfully that they present serious additional risks), and are able to autonomously solve real-world paid software engineering tasks, it does not seem implausible that models in the GPT-4.5 to GPT-5 class will be able to autonomously perform cyber attacks or pose significant challenges for biosecurity. A tentative and adjustable threshold of 10^26 FLOP thus seems warranted from a precautionary principle. The height for this classification threshold is, of course, up for debate and could be further refined through inclusive dialogues with industry, academia and civil society. The Commission should also leave open the possibility to adjust the compute threshold upwards, in case of limited algorithmic progress and fast-growing societal resilience to very capable models.

Requirements for the additional GPAI tier would have to be weighed on axes like risk-reduction, regulatory burden, and feasibility. This will necessitate thoughtful analysis of how requirements can backfire or fail to prevent accidents. Possible requirements could include—but are not limited to:

- Pre-development third-party evaluations to assess whether the AI company (and possible third-party cloud service provider) have taken sufficient infosecurity measures.

- Pre-development third-party evaluations to assess whether the AI company's alignment procedures are sufficient given the expected model capabilities.

- On-premise involvement by the AI Office during training to evaluate model checkpoints on dangerous capabilities and to perform extensive pre-deployment evaluations.

It may seem strange to already begin extending the AI Act before the Act has even entered into force. However, the pace of progress in AI calls for unconventional approaches. If current compute trends continue, we will likely see models trained using more than 10^26 FLOP be released before the AI Act's requirements for GPAI even start taking effect (these will only apply 12 months after the AI Act enters into force, in July 2025). Given that negotiations and implementation take time, the best moment to start preparation for a third GPAI tier is today. The codes of practice for GPAI models form an excellent temporary opportunity to add additional pre-development guardrails for very large (e.g. more than 10^26 FLOP) GPAI models until a third tier can be solidified in the main regulation.

RECOMMENDATION 3

Scale and prioritize enforcement funding based on compute trends

The AI Office will fulfill a crucial role in safeguarding the EU from systemic risks posed by GPAI models. After all, the AI Act will only be effective insofar as it can be adequately implemented and enforced. In order to do so the AI Office will likely have to increase its resolve to hire top-notch technical experts and prioritize enforcement of the most risky models.

As is reflected in the AI Act, compute can be used as a proxy for a model's potential to introduce systemic risk. The more compute a model is trained with, the more advanced its capabilities will typically be and thus the larger the potential user base and the risk of unwanted consequences. If implementation and enforcement capacity is limited - which may very well be the case with the AI Office still under construction - prioritizing efforts based on levels of training compute may thus be required to safeguard EU citizens from the largest risks. This does not only pertain to the different treatment of 'regular' GPAI models and GPAI models with systemic risk: as compute budgets are rapidly rising, we can also expect significant variation in risk-potential within the class of models trained on more than 10^25 FLOP. To adequately enforce the AI Act in times of personnel constraints, it may be necessary to spend the bulk of the AI Office's evaluation capacity on only a handful of models, only shallowly checking for potential infringements by providers of less-compute intensive models.

Recent hardware orders suggest that the 10 leading AI labs could pass the AI Act's compute threshold in 2024

Of course, this situation is far from ideal: less compute-intensive models can still cause large-scale harm, especially when embedded in products that are used by millions of people. The AI Office is aiming to hire approximately 100 employees by the end of 2025, of which only a limited number will be technology specialists (the EU's experience with the GDPR shows that having sufficient technology specialists is crucial for proper enforcement) . Although a great start, the AI Office's capacity could become insufficient quite rapidly. Recent AI chip orders suggest that most of the 10 leading AI companies will already pass the 10^25 FLOP threshold in 2024, possibly for multiple models per company[16]. If we assume the number of above-threshold models to grow in line with compute trends, we could see the release of more than 100 GPAI models with systemic risk in 2026[17]. If we assume the AI Office hires 20 technology specialists (the Office will also hire administrative assistants, policy experts and legal experts)[18], and that at least half

[16] NVIDIA is currently making roughly 20 billion USD on datacenter revenue per quarter. Assuming an average selling price of 20,000 USD per H100 GPU, this suggests they are selling some 1 million AI chips per quarter, or 4 million on a yearly basis (this estimate seems conservative as most analysts predict NVIDIA's revenue growth to continue throughout 2024 and 2025). GPT-4 - the first GPAI model with systemic risk - required less than 10,000 H100-equivalents to train. 4 million H100's could thus theoretically be used to train 400 GPAI models with systemic risk. In reality, this number will be much smaller since most AI companies are pursuing increasingly large training runs and are also using their chips for experimentation and inference. Nevertheless, reports indicate that at least 12 companies have access to enough AI chips to train GPT-4 level models in 2024 and some of them may release multiple >10^25 FLOP models.

[17] The actual number could turn out significantly smaller if winner-takes-all dynamics are so large that funding for new players dries out, or if developers focus all their attention on the release of 1 very capable model per year. Betting on such an outcome, does not seem prudent, however.

of those will not be working on enforcement and evaluation (they are for instance tasked with shaping the codes of practice, or building the necessary hardware infrastructure for model evaluations), this could leave 10 technology specialists to oversee compliance of 100 GPAI models and to conduct evaluations and investigations when needed. Although it is too soon to tell with confidence how much work will result from the release of a single GPAI model with systemic risk, these numbers do not seem reassuring.

Of course, the above capacity projections are very rough and may not match the envisaged proportion of technical experts. The Commission should hence carry out or commission a more detailed projection that is grounded in both compute trends and internal data. Capacity projections can further be updated once the AI Office has built more experience carrying out its tasks such that capacity needs per GPAI model release become clearer.

Whatever the outcome of such exercise, attracting sufficient skilled personnel may prove challenging. While technical staffers at the large AI companies typically make upwards of 400,000 USD a year, yearly compensation for similar roles at the AI Office is currently limited to only 50,000-60,000 euros a year. Besides the obvious - but politically difficult - solution of increasing compensation, capacity constraints can also be overcome by working together more closely with academia and civil society, for instance to help shape the codes of practice, or to develop methodologies and benchmarks.

[18] Within the European DPA's only some 10% of employees are tech specialists, so this would already constitute a relatively large fraction.

RECOMMENDATION 4

Create a dedicated foresight unit within the AI Office

The AI Office is tasked with '[monitoring the evolution of AI markets and technologies](#)'. Future-proof regulation and enforcement, however, not only necessitates monitoring previous developments, but also looking out for potential future developments. The European Commission should thus create a dedicated AI foresight unit within the AI Office to help EU policy makers skate where the puck is going. Studying (effective) compute trends in AI would enable this unit to discern several quantitative scenarios of future training compute budgets and inference capacities. Building on those scenarios, the foresight unit could work together with academia and civil society to map out what types of capabilities and accompanying risks might arise in the coming years.

If data center growth continues along current trends, it could place significant stress on the EU's renewable energy goals

The EU had to compensate for a lack of foresight by significantly adapting the AI Act after GPAI models started to shake the world. If progress in AI continues at current pace, the AI Act could be out of date before the EU has even properly attempted to enforce it. Understanding compute's central role in AI development will be crucial to create future-proof updates and to prevent haphazard adjustments as much as possible. For instance, it seems worth already mapping out ways the EU AI Act may fall short when compute growth enables AI agents that can autonomously perform complex tasks in the digital domain. Agent-structures can be relatively easily built on top of large GPAI-models, but can add many new capabilities (like autonomous decision making, or long-term action taking). AI Agents could therefore pose new challenges for governance throughout the entire value chain and may necessitate moving various requirements upstream to the core foundation model providers.

Knowledge of compute growth can also help policy makers anticipate challenges in different sectors. For instance, if data center growth continues along current trends, this could place significant stress on the EU's renewable energy goals. Simultaneously, it could make concerns that spurred the EU Chips Act even more pressing. This tension between sustainability on the one hand and competitiveness on the other will likely require the EU to make difficult trade-offs. To manage such trade-offs requires understanding the future implications of different choices - in other words, it requires foresight.

The AI Office seems uniquely positioned to house such foresight work given the expertise it will build in evaluating risks from GPAI models and its proximity to the large AI companies. By further leveraging collaborations with academia and civil society, the AI Office could conduct state-of-the-art foresight work that forms the basis of future-proof AI governance. Note that it may be too late or too burdensome to add this envisaged foresight task to to the AI Office's formal responsibilities. In that case, the Commision should look for practical workarounds, such as embedding foresight work in the AI Office's monitoring efforts.

RECOMMENDATION 5

Implement a multilateral compute oversight system

There exists very limited public information on how many AI chips are sold to different companies, where the largest clusters are located and which providers are currently pursuing large training runs. As a result, EU policy makers are operating in the dark. If the EU does not know how much compute is out there and how it is distributed, it will face tremendous difficulties making the right investment decisions, preparering future extensions of the AI Act, anticipating future AI developments and enforcing the AI Act. Other jurisdictions face exactly the same issues. To overcome the problems, the EU needs to start international dialogues to put in place a multilateral compute oversight system. By no means should participating countries of such a system try to keep track of the location of every individual AI chip (this would no doubt require serious privacy violations). Instead, necessary information-flow could be limited to the reporting requirements that are currently being developed under the US Executive Order 14110:

- The location and owner of any large AI cluster (theoretical maximum of >10^20 FLOP/s and >100 gbit/s networking)
- The size, location, provider and compute provider of any planned or ongoing 'very large' training runs (>10^26 FLOP)

To prevent the spread of sensitive non-public information as much as possible, these requirements could apply within each participating jurisdiction, and collaborating countries could commit to sharing high-level decision-relevant information with each other. Establishing such an oversight system will be a politically challenging task and will require open dialogue and systematic coordination with third countries and target companies across the EU and internationally. However, it is also a task the importance of which can hardly be overstated.

To increase the chances of success, the EU can start out by shaping a bilateral US-EU compute oversight system. As most of the leading AI companies and cloud providers are located in the US, a bilateral compute oversight system can already reap most of the benefits that come from proper compute monitoring. After a successful launch of a bilateral agreement, the system can be extended to other countries. To make the burden for the US to participate in such a system as low as possible, the EU could propose to extend the core reporting obligations of the US Executive Order 14410 (EO) to its own jurisdictions.
Note that the model reporting thresholds of the EO are higher than the GPAI thresholds in the AI Act (10^26 FLOP vs 10^25 FLOP). This does not mean the EU should adjust the GPAI classification threshold of the AI Act - in fact, the EO's model reporting requirements add quite naturally to the AI Act as they do not exclude training runs for the 'sole purpose of research, development and

prototyping activities'. The more compute-intensive such 'internal' training runs are, the more risky they become. It thus makes sense to extend reporting requirements to all large training runs above a certain compute-intensity. Furthermore, the AI Act does not yet yield reporting requirements for owners of large compute clusters. Given that these compute providers play a key role in maintaining proper security of large GPAI models during training, involving them in the governance of advanced AI models seems highly useful.

A multilateral compute oversight system requires harmonized compute accounting standards: clear and detailed instructions on how to measure cluster sizes and training run sizes

The EO's reporting requirements only target very large compute clusters (>50,000 NVIDIA H100's) and multi-week training runs on such clusters. Currently, clusters of this size do not yet exist in the EU (they are, however, being built in the US). With compute budgets rapidly increasing, however, it is a matter of time until many of such clusters will be operational in multiple countries. It may seem that this results in hard-to-enforce reporting requirements. However, here the heavy market concentration of the AI supply chain plays into the hands of regulators: there are very few companies that have the investment capacity and know-how to build (and rent out) multi-thousand AI-chip clusters. Typically such build-outs are reserved to a handful of big-tech companies such as Apple, Meta and Bytedance that require giant clusters for their own services, and to cloud service providers like Microsoft, Amazon, Google, Oracle and Tencent. Moreover, there already exist concrete proposals on how to involve such players via so-called Know-Your-Customer checks to improve compliance with regulation and to provide regulators with essential monitoring data in privacy-preserving ways.

A multilateral compute oversight system requires harmonized compute accounting standards: clear and detailed instructions on how to measure cluster sizes and training run sizes. Industry, academia, and civil society have already proposed core elements of such compute accounting standards. These approaches will have to be further refined and tested in the real world. The EU is uniquely positioned to experiment with compute accounting standards through the EuroHPC Joint Undertaking. The recent push to open up European supercomputers to AI startups provides perfect timing to explore how compute reporting can work in practice without overburdening the cluster operators or risking leakage of proprietary data. In collaboration with EuroHPC, the AI Office could develop harmonized compute accounting standards that describe how owners of large clusters should report their hardware resources and training run characteristics. These standards can subsequently form the foundation of the multilateral compute oversight system. They can also be embedded in the delegated acts that amend the reporting requirements of the AI Act's Annex IXa and can inform the Codes of Practice for GPAI models with systemic risk, killing three birds with one stone.

Eventually, cooperation with other jurisdictions on a compute oversight system with proper compute accounting standards could result in a joint database with high-level information on compute clusters and training runs that is adequately protected and can only be accessed by the countries' respective enforcement agencies.

Such a joint database can be an attractive solution for participating members as is for instance shown by the success of the International Methane Emissions Observatory. Lessons on how to set up and properly protect this joint database can be taken from the EU's effort on topics such as child protection, and disinformation. The creation of a joint compute oversight database will no doubt run into all kinds of questions surrounding privacy and security that have to be ironed out through international dialogue. The upcoming AI Safety Summits in South Korea and France make for excellent opportunities to start this conversation.

Conclusion

The exponential increase of compute has been the predominant driving force behind recent progress in advanced AI.

Scaling laws and industry investments suggest that compute will likely play a similar role in the remainder of this decade.

Within the next few years, compute growth could enable new AI capabilities that revitalize European productivity, but could also disrupt job markets, introduce new threats in cyber and biowarfare, or lead to large-scale accidents. Given the unique role compute plays in the AI supply chain, EU policymakers should urgently increase capacity and expertise on compute governance as a way to steer clear of such risks. With the AI Act on the cusp of entering into force, now is the time to start harnessing compute as a policy lever for democratic oversight and enforcement. To stay ahead of the AI revolution, the EU should consider:

1. **Investing in compute** in a targeted way to promote our understanding and control of advanced AI models, and to train specialized AI systems that can help tackle large societal issues in e.g. medicine, energy and climate science.

2. **Preparing the addition of a third tier of regulation** for GPAI models with severe systemic risk, using a classification mechanism that is partially based on training compute.

3. **Strengthening the AI Office's resolve** to prioritize evaluation of the most compute-intensive GPAI models in case of (temporarily) limited personnel capacity and to increase capacity in line with compute trends.

4. **Adding a dedicated foresight unit** to the AI Office that helps anticipate future AI policy challenges based on (effective) compute trends.

5. **Leading the way** on establishing a multilateral compute oversight system that can aid each of the previous 4 efforts.

→ About ICFG

The International Center for Future Generations (ICFG) is an independent think-and-do tank dedicated to shaping a future where decision-makers anticipate and responsibly govern the societal impacts of rapid technological change, ensuring that emerging technologies are harnessed to serve the best interests of humanity.

Our experts supply rigorous, independent, policy-oriented research and analysis that connects the dots between the frontiers of emerging technology development.

→ About the author

Daan Juijn is an analyst at ICFG, responsible for foresight work within the focus areas of Advanced AI, Climate Interventions and Quantum Computing. He specializes in quantitative scenario analysis and uses his interdisciplinary perspective to map out possible future interactions between emerging technologies.

Image credits: Cover image generated with AI Shutterstock (reference to Giorgio de Chirico)

www.icfg.eu

International
Center for
Future
Generations ICFG