Future of Compute Review - Call for Evidence

Response by the Centre for the Governance of AI

August 5, 2022

Lennart Heim
Research Scholar
Centre for the Governance of AI
lennart.heim@governance.ai

Markus Anderljung
Head of Policy
Centre for the Governance of AI
markus.anderljung@governance.ai

About the Centre for the Governance of AI (GovAI)

The Centre for the Governance of AI (GovAI) is a nonprofit based in Oxford, UK. It was founded in 2018, initially as part of the Future of Humanity Institute at the University of Oxford, before becoming an independent research organisation in 2021. GovAI's mission is to build a global research community, dedicated to helping humanity navigate the transition to a world with advanced AI. More information at governance.ai.

Lennart Heim focuses on compute governance. His research interests include the role of compute in the AI production function, the compute supply chain, forecasting emerging technologies, and the security of AI systems. He's also a member of the OECD.AI Expert Group on AI Compute and Climate. He has a background in Computer Engineering.

Markus Anderljung's work aims to identify and improve AI governance policy recommendations. His research focuses on the potential global diffusion of EU AI policy, the regulation of AI, surveys of AI researchers, compute governance, and responsible research norms in AI.

Key Recommendations

We recommend that the Future of Compute Review:

1. Considers how best to support AI research in academia
2. Encourages the prevention of misuse of AI capabilities while still preserving safe access, via structured access to ML systems
3. Aims to enable research to scrutinise and understand high-compute and impactful models by researchers outside AI labs
4. Considers a tiered access approach to compute provision, where access to larger amounts of compute comes with additional review requirements, if the Review recommends the creation of a compute fund
5. Recommends cooperating with allies in securing and stabilising the semiconductor supply chain

1. Background

We welcome the opportunity to respond to the [Future of Compute Review's call for evidence] and look forward to future opportunities to provide additional input. We offer the following submission for your consideration.

Our response focuses on the future of compute used for Artificial Intelligence (AI). In particular, we emphasise the risks posed by advanced AI systems that are enabled by access to large amounts of compute.

Compute is a necessary resource for Machine Learning (ML) systems. Next to data and algorithmic advances, compute is a fundamental driver of advances in AI. The compute required for training notable ML systems has been doubling every six months — growing by a factor of 55 million over the last 12 years.[1] Cutting-edge research in ML has become dependent on access to large amounts of compute and the expertise to leverage them.

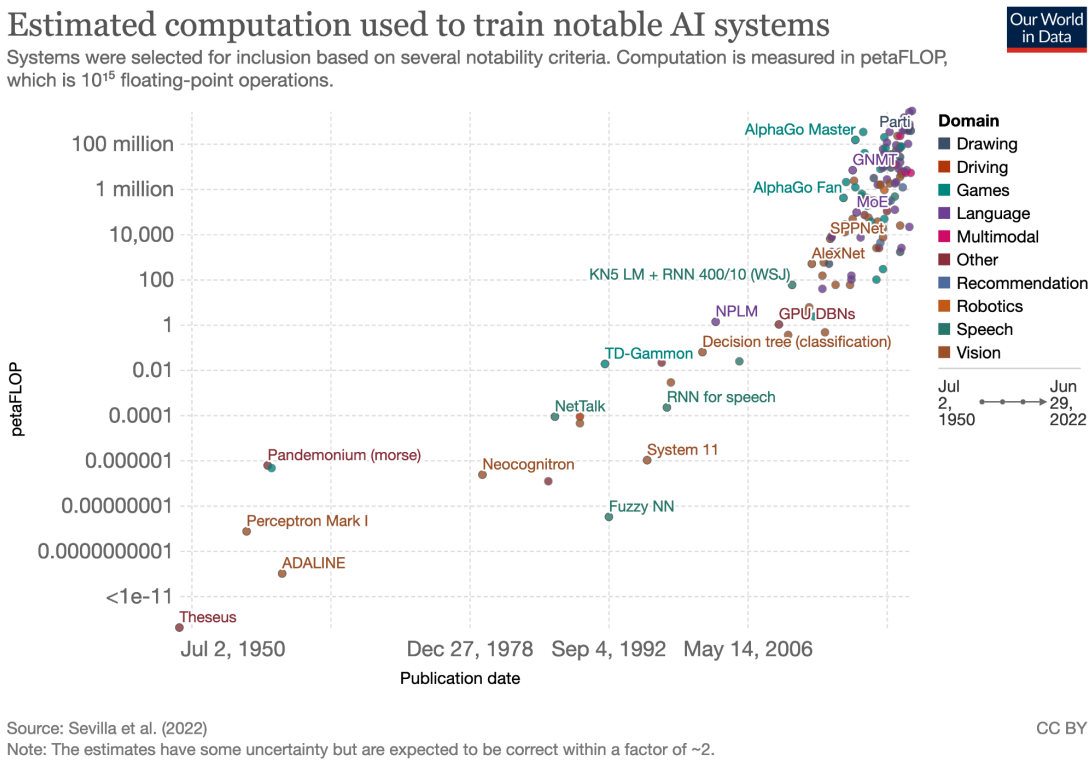

Figure 1: Estimated compute used for training notable AI systems over the last 70 years. Since 2010 the compute used for training has been doubling every six months (Sevilla et al. 2022).

[1] Sevilla et al. 2022 "Compute Trends Across Three Eras of Machine Learning"

Over the next few decades, we believe that the amount of compute used to train an AI system will become an increasingly useful metric when considering what responsible AI practices are appropriate. Firstly, the performance of ML models tends to scale with compute.[2] State-of-the-art models across domains, such as PaLM (Google), AlphaFold (DeepMind), and GPT-3 (OpenAI), tend to outperform competing models in part by using vastly more compute.[3] For example, it took more than 64 days across thousands of chips to train PaLM with a cloud computing equivalent cost of $9M to $23M.[4] While the performance of a system also scales with the amount and quality of data,[5] there are no agreed-upon metrics of data quality that could be used for this purpose.

Secondly, the performance of an AI system is a useful proxy of its potential impact, both positive and negative. As many scholars have argued, AI systems can pose a variety of risks. As AI systems become more capable, we should expect these risks to increase, with their potential to be misused[6] or cause accidental harm.[7] In short, the more capable the system, the more uses it can be put to, and the more important it is that it is developed and deployed responsibly.

Throughout our provided evidence, we continuously refer to the following key recommendations which help to address these risks.

Recommendations

(1) Consider how best to support AI research in academia

One of the key trends in AI research over the last decade is its growing need for computational resources. Private AI labs are producing an increasing share of these high-compute SOTA AI models,[8] leading many to worry about a growing compute divide between academia and the private sector.[9] Importantly, this is not just a divide with regards

---

[2] Kaplan et al. 2020 "Scaling Laws for Neural Language Models"; Hofman et al. 2022 "Training Compute-Optimal Large Language Models"
[3] Sevilla et al. 2022 "Compute Trends Across Three Eras of Machine Learning"
[4] Chowdhery et al. 2022 "PaLM: Scaling Language Modeling with Pathways"; Heim 2022 "Estimating PaLM's training cost"
[5] Model size (number of parameters) and number of data samples are linearly correlated with the amount of compute. However, it's independent of the quality of data.
[6] Brundage et al. 2018 "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation."
[7] Zwetsloot & Dafoe 2019 "Thinking About Risk From AI: Accidents, Misuse, and Structure"
[8] According to Sevilla et al., 2022, every AI system that has set a new record for compute consumption since 2016 has been produced by a private lab.
[9] Ahmed & Wahed 2022 "The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research"; Ganguli et al. 2022 "Predictability and Surprise in Large Generative Models"

to computational resources required to train SOTA models, but also to use and run experiments on them.[10]

Partly in response to these concerns, there have been calls for national compute provision in a number of countries. For instance, in the United States, the National AI Research Resource (NAIRR) has recently been proposed.[11] The NAIRR would help provide academic researchers with access to compute, by either operating its own compute clusters or distributing credits that can be used to buy compute from other providers.[12] It would also further support academic researchers by granting them access to data, including certain government-held datasets. If the Future of Compute Review looks into similar options to address the compute divide, we encourage it to take a broad view, asking: "How can AI research by academics best be supported?" and not simply "Should the UK provide compute to academic AI researchers?".

In particular, we suggest exploring the following questions: Is providing compute directly preferable to giving researchers money that they can spend as they see fit, be it on talent, compute or datasets? One might expect researchers themselves to make better fund allocation decisions than a central planner.

Potential reasons in favour of providing researchers with compute resources rather than funding that we encourage the compute review to explore include: Researchers may not be able to invest sufficiently in compute, e.g. because their university procurement processes make it difficult. Further, academic researchers may be more interested in theoretical questions, rather than producing state-of-the-art high-compute systems.

There may also be large economies of scale that make it preferable for the government or some central actor to procure computing infrastructure on behalf of the whole nation's researchers. For example, a national compute fund could negotiate better prices with cloud compute providers than individual researchers could. Should there be sufficient expertise and compute demand, further cost reductions may be possible via setting up dedicated cloud computing infrastructure. Are these economies of scale large enough to justify an in-kind rather than monetary subsidy to academic researchers?

[10] For example, a recently released open-source 176B Multilingual Language Model "BLOOM" requires more than 350GB of memory just for execution. A researcher's single computer is not sufficient for exploring and analysing this ML system. Whole systems of AI accelerators, for example GPUs, are required. As an example, an NVIDIA DGX A100 640GB system would be sufficient but it cost more than $199,000 at its release.
[11] Etchemendy & Li 2020 "National Research Cloud: Ensuring the Continuation of American Innovation"
[12] Ho et al. 2021 "Building a National AI Research Resource"

(2) Encourage the prevention of misuse of AI capabilities while still preserving safe access, via structured access to ML systems

As AI systems become more capable, it will be more possible to use them to cause harm.[13] As an example, a recent paper reports how a group of scientists used their AI system designed for drug discovery to identify potential new toxins by having the system find molecules with maximum rather than minimum toxicity.[14] The paper reports the scientists' model identifying multiple new toxins, predicted to be more toxic than publicly known chemical warfare agents.

To prevent such misuse, we suggest that the Review looks into steps to encourage the providers of cutting-edge AI systems with significant misuse potential in the UK (both private and governmental) to make some of their models available via structured access approaches.[15] This entails giving users limited access to a model e.g. via an application programming interface (API), allowing them to use the model (hosted e.g. on the cloud) without running it on their own systems. By putting the model behind an API, it is possible to monitor what the system is being used for (e.g. a provider of a text-generation model can see if a user is outputting reams of political content) and to restrict certain uses (e.g. one could ensure a drug discovery model does not output any result where the predicted toxicity is above some specified level). If a model is open-sourced, on the other hand, it is very difficult to ensure monitor or control what it is used for.

If the compute review decides to set up a UK compute fund, we recommend that it includes the necessary infrastructure to host models behind an API.[16] This would be an important tool in ensuring that models trained using compute provided by the government are not misused.

(3) Enable research to scrutinise and understand high-compute and impactful models by researchers outside AI labs

We are particularly worried about the compute divide leading to less external expert scrutiny of and research to understand models trained with significant amounts of compute or that are particularly impactful. As AI systems become more capable and make more important decisions in people's lives, our need to understand their impacts and what produces certain model behaviour increases. We will increasingly need research into how AI systems might unexpectedly fail or cause harm, and what can be done to mitigate these issues. This includes work into models' robustness,[17] fairness, and interpretability. We may also need increasing scrutiny of particularly impactful models, investigating how they may

[13] Brundage et al. 2018 "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation."
[14] Urbina et al. 2022 "Dual use of artificial-intelligence-powered drug discovery"
[15] Shevlane 2022 "Sharing Powerful AI Models"
[16] Anderljung et al. 2022 "Compute Funds and Pre-trained Models"
[17] Rudner & Toner 2021 "Key Concepts in AI Safety: Robustness and Adversarial Examples"

be causing issues. As such, we encourage the Compute Review to consider how research to scrutinise and understand high-compute and particularly impactful models can be conducted outside AI labs.

If the UK chooses to develop a compute fund – providing academic researchers with access to computational resources, similar to the US proposed National AI Research Resource – we would recommend that the compute fund creates an API which can host pre-trained models, developed with compute fund resources, private labs, academics, or government bodies.[18] That API could be used to allow external scrutiny of societally impactful models, e.g. systems used by the public sector to identify suspected tax fraud.

It is also worth exploring how algorithms used to moderate and curate content on social media platforms could receive greater outside scrutiny. For example, Twitter has recently created an API that allows outside researchers to scrutinise platform data in a privacy-preserving manner.[19] Building on this, could researchers or government officials be given API-access to social media companies' recommendation algorithms to run experiments on whether they e.g. tend to increase polarisation.

(4) Consider a tiered access approach to compute provision, where access to larger amounts of compute comes with additional review requirements, if the Review recommends the creation of a compute fund

If the Review recommends the creation of a compute fund, we propose a tiered access approach to compute provision – where access to larger amounts of compute comes with additional review requirements regarding responsible AI practices – to address the risks stemming from AI models developed with (subsidised) compute.

Since compute is a finite, rivalrous resource, the UK government will have to make difficult decisions about how it is allocated. Such decisions should be based on many factors, including scientific merit and practicability. Importantly, it should also be based on the extent to which the researchers adhere to responsible AI practices, e.g. foreseeing and preventing potential risks the model could impose. The more compute a project is allocated, we argue, the greater care should be taken by the government and the researchers to reduce risks and spread the benefits of the system.[20]

More generally, we expect that regulatory burdens or responsible AI requirements should, among other things, start to scale with the capabilities of the relevant systems in coming decades. More capable systems will tend to offer higher potential benefits and pose larger risks. As the amount of the compute used to train a system is a useful proxy for its

[18] More details in Anderljung et al. 2022 "Compute Funds and Pre-trained Models"
[19] Twitter 2022 "Investing in privacy enhancing tech to advance transparency in ML"
[20] We expand on this recommendation in our response to the US National Science Foundation's RFI on the proposed National AI Research Resource: Heim & Anderljung 2022 "Comments on the interim report of the National Artificial Intelligence Research Resource Task Force"

performance, we expect that the amount of compute could become a particularly important factor that determines the regulatory burdens imposed on future systems.

(5) Cooperate with allies in securing and stabilising the semiconductor supply chain

We encourage the UK to explore strategic partnerships with their allies, e.g. the US, the EU, Taiwan, and NATO, to secure and stabilise a steady supply of semiconductors.

Compute is enabled by the complex and concentrated semiconductor supply chain, which has been subject to disruptions over the last years.[21] In response to these shocks, many nations have explored initiatives to secure the supply chain, e.g., CHIPS act in the US[22] and the European Chips Act.[23] We encourage the UK to explore the role of the UK in the semiconductor supply chain and collaborate with its allies to secure the supply.

2. Our response to the questions

2.1 Users

1. What are the compute needs of UK users?

Compute is a key input to ML systems. Next to data, and algorithmic advances, increasing amounts of compute has been an important driver of recent advances in AI. The compute required for training notable ML systems has been doubling every six months — growing by a factor of 55 million over the last 12 years.[24] Cutting-edge research in ML has become synonymous with access to large compute budgets or computing clusters, and the expertise to leverage them.

This growth in training compute is unprecedented and largely enabled by increased spending on compute — buying and renting more specialised AI chips for longer periods. In contrast, the performance of the best-performing high-performance computer has only grown by a factor of 245x in the same period (compared to the factor of 50M in growth in required training compute).[25] Also, the price performance of GPUs, the predominant type of hardware used to train state-of-the-art ML systems, has *only* doubled every 2.5 years (in contrast to the 6-months doubling time of required training compute).[26]

---

[21] Khan 2021 "The Semiconductor Supply Chain"
[22] CHIPS for America Act & FABS Act - Semiconductor Industry Association
[23] European Chips Act: Communication, Regulation, Joint Undertaking and Recommendation
[24] Sevilla et al. 2022 "Compute Trends Across Three Eras of Machine Learning"
[25] Referring to the performance of the leading high-performance computer of the TOP500 from November 2009 to 2021.
[26] Epoch 2022 "Trends in GPU price-performance"

2. How do you expect demand for compute in the UK to change over the next decade?

As outlined before, the current trend of compute demand of AI systems is unprecedented. This is highlighted by the private AI sector either massively investing in their own compute resources[27] or joining strategic partnerships for access to compute[28] in recent years. The fact that many leading AI labs are also leading cloud providers underlines this.[29]

While we do not expect this trend in increased investments in compute to continue indefinitely due to the enormous compute costs surpassing economic limits soon[30], many sectors, such as academia and small- and medium-sized companies, are currently not able to participate in frontier high-compute AI research. For those sectors to catch up, they require sufficient access to compute and the connected enablers, such as a skilled workforce, to train frontier models. They could also be provided access to high-compute models developed by other actors and associated computing infrastructure, e.g. via API access as described above.

3. Should the government be trying to stimulate demand for compute, and why?

See our recommendation (1).

2.2 Infrastructure

4. How do you expect the compute provision that is available to UK users to change over the next decade?

We expect a trend of AI systems being offered more widely as a service — becoming a more dominant part of the software as a service (SaaS) industry.

For example, we predict that an increasing portion of important AI research and development will make use of large pre-trained models that are accessible only through APIs. In this paradigm, pre-trained models would play a central role in the AI ecosystem. A large portion of SOTA models would be developed by fine-tuning[31] and otherwise adapting these models to particular tasks. Commercial considerations and misuse concerns would also frequently prevent developers from granting others access to their pre-trained models,

[27] See for example Meta's recent announcement: Meta's cutting-edge AI supercomputer for AI research
[28] OpenAI & Microsoft 2019
[29] Amazon with AWS; Google with Google Cloud
[30] Lohn & Musser 2022 "AI and Compute: How Much Longer Can Computing Power Drive Artificial Intelligence Progress?"; Carey 2019 "Interpreting AI compute trends"
[31] Fine-tuning describes the process of improving the performance of a pre-trained model on a specific task by training it on a task-related dataset.

except through APIs. Though we are still far from being in this paradigm, there are some early indications of a trend that we discuss in Anderljung et al. 2022.

5. How should the government incentivise the supply of compute?

No response.

6. What ownership and operational models could best meet the needs of compute users?

We refer to our recommendation (1).

7. What are the risks of the increasingly widespread use of compute, and how can they best be mitigated?

Our background has focused on the risks posed by advanced AI systems, and we propose recommendations (1), (2), (3), and (4) to address them.

8. How can the government most effectively intervene in the compute market to help to mitigate the environmental impact of this technology?

Use the location flexibility of compute and use already existing *green* datacentres. Newly established datacentres should be placed close to renewable energies, while also attending to latency requirements, which does not necessarily need to be within the UK.

2.3 Access and enablers

9. How can the government help to increase access to compute across user groups?

See our recommendation (1), where we discuss this and propose questions to investigate.

10. What are the key issues that prevent UK users accessing the compute supply?

Lennart Heim, one of the authors of the response, has co-authored a report for the OECD expert group on AI Compute and Climate[32] which comments on this question. It will be made publicly available later this year.

11. What public procurement approaches could best meet compute users' current and future needs?

No response

[32] Expert Group on Compute & Climate - OECD.AI

2.4 International approaches to compute

12. How does the UK compare internationally in relation to compute? Does the UK have, or should it develop, specific strengths?

No response.

Disclaimer

We are responding on behalf of The Centre for the Governance of AI.

The Centre for the Governance of AI has no financial or other links to any company operating in this sector. None of our research reports has received commercial funding from companies operating in the compute sector.

Lennart Heim reports no conflict of interest. He is a member of the OECD.AI expert group on AI Compute & Climate but responds to this call as an employee of The Centre for the Governance of AI.

Markus Anderljung reports no conflict of interest. He is currently on a part-time secondment to the Brexit Opportunities Unit in the Cabinet Office, but responds to this request in his capacity as an employee of the Centre for the Governance of AI.