Future of compute review - submission of evidence

Prepared by:
Dr Jess Whittlestone, Centre for Long-Term Resilience (CLTR)
Dr Shahar Avin, Centre for the Study of Existential Risk (CSER), University of Cambridge
Katherine Collins, Computational and Biological Learning Lab (CBL), University of Cambridge
Jack Clark, Anthropic PBC
Jared Mueller, Anthropic PBC

Jess Whittlestone is responding on behalf of the Centre for Long-Term Resilience. Other authors are responding in a personal capacity and the submission does not necessarily represent the views of their affiliated organisations.

Shahar Avin contracts with Viz.ai, a medical AI company, and is an advisor to the Machine Intelligence Garage which supports UK startups.
Jack Clark and Jared Mueller both work at Anthropic, a compute intensive AI startup based in California.
Jack Clark is on the advisory board of the UK Centre for Data Ethics and Innovation (CDEI)

Contents

Executive summary

We are a group of experts in AI governance with experience across academia, industry, and government. Our evidence submission is framed around the question of **how the UK's compute strategy can help achieve the goals of the National AI Strategy**: investing in the long-term needs of the AI ecosystem; ensuring AI benefits all; and governing AI effectively.

Access to increasingly large amounts of computing power has been a key driver of AI progress in recent years. In order to leverage the benefits of this progress for society and the economy, the UK government must effectively and proactively manage risks. Compute-intensive AI progress is particularly likely to lead to more systemic, high-stakes, and difficult to anticipate impacts, requiring anticipatory governance approaches to manage.

By taking more proactive measures to understand and influence how large scale-compute is used to drive AI progress, the government can more effectively ensure that AI benefits all of society.

We make two specific recommendations for how the UK can do this, which we suggest this review should explore:

1. Our first recommendation is that **the review should explore ways to increase compute capacity for academia**, especially researchers working on beneficial AI applications, AI safety and security research (relevant to questions 1-3 and 9-11).

   Given the importance of compute for AI progress, which groups have access to large-scale compute will determine the interests and incentives that shape AI development.

   There is currently a large disparity between the computational resources available to AI researchers in academia and industry, and evidence of substantial latent demand for compute among academic researchers.

   Increasing compute access for academia would strengthen the UK's AI ecosystem while improving the scrutiny and accountability of commercial research and shifting incentives towards longer-term benefits for society.

   We also outline additional measures which could support academics' ability to contribute to compute-intensive AI research, including providing access to trained models via APIs.

2. Our second recommendation is that **the review should consider how the UK government could monitor and govern compute usage** in AI development more broadly (relevant to questions 7 and 12).

   In order to effectively govern AI, the UK government needs better information about potential harms, and more effective tools for intervening to prevent harm.

> Compute is a strategic leverage point which can help with both these challenges: providing information about potentially high-risk uses of AI, and a lever by which the government can practically shape AI progress.
>
> Practically, compute is much more easily monitored and governed than other inputs such as data, algorithms, and talent, due to being relatively easy to quantify, and being highly centralised in its supply and use.
>
> We outline a 'tiered' approach to compute-indexed AI monitoring and governance, and explain how this could strongly support the government's aim of establishing a pro-innovation approach to regulating AI, as well as accurately assessing long-term AI safety and risks.

We first provide some background on compute as an important policy lever for shaping beneficial AI development, before outlining the two recommendations above in more detail.

Background: compute as a lever for shaping beneficial AI development

Access to increasingly large amounts of computing power has been a key driver of AI progress in the past few years (Kaplan et al 2020; Sevilla et al 2022). In particular, large-scale compute has recently enabled dramatic progress in "foundation models" - including language and image generation models - which are showing impressively general performance across a wide range of tasks within these domains (Bommasani et al, 2022). There are good reasons to expect this trend of increasingly compute-intensive AI models to continue, leading to more capable and general-purpose AI capabilities (Amodei and Hernandez 2018; Kaplan et al 2020).

While this rapid AI progress has potential to benefit society and the economy, it also poses risks. In recent years, we have already begun to see harms arising from narrow AI systems deployed in specific domains, including self-driving car accidents (Nelson 2015), and bias in algorithms used in a range of contexts from healthcare (Wiens et al 2020) to criminal justice (Angwin et al 2016) and loan approval (Bartlett et al 2022). Researchers in academia and industry have begun to document the various harms that have arisen or could easily arise from compute-intensive language models, including discrimination, toxicity, misinformation, malicious use, and the wider societal consequences of loss of employment (Ganguli et al 2022, Weidinger et al, 2021, Bender, Gebru et al. 2021; Bommasani et al, 2022).

Managing these risks effectively, and building strong public trust in the UK's AI ecosystem, will be central to the UK's vision of becoming a global AI superpower.

To some extent, it is possible to address these harms on an individual basis and via sector-specific regulation. But **compute-intensive AI progress is likely to lead to more systemic, high-stakes, and difficult to anticipate impacts, making this approach increasingly challenging**. Compute-intensive AI systems tend to have emergent capabilities that are difficult to predict in advance (Ganguli et al 2022), which can lead to unexpected harms. And as we are

already beginning to see with language models, the increasing availability of compute is making it easier to develop large-scale AI models with impressive capabilities and applications across a huge range of domains. This could precipitate much more extreme societal impacts than those we are already dealing with today: for example, the use of increasingly advanced AI systems by militaries could make conflict or escalation more likely; increasingly capable AI systems used for content generation and persuasion could undermine democratic debate; and the unequal distribution of harms and benefits from AI progress could lead to more extreme power concentration and inequality in society (Brundage, Avin et al. 2018; Seger et al 2020; Clarke and Whittlestone 2022; Whittlestone and Clarke 2022).

In addition to addressing specific harms as they arise, there is therefore also a growing need for the UK government to explore approaches to **anticipatory governance**: how can the government proactively shape AI innovation in ways that prevent harms from occurring, therefore helping unlock the benefits of AI for society and the economy? Our overarching suggestion in this evidence submission is that compute access and usage may be an especially important policy lever for the UK government in this regard: **by taking more proactive measures to shape how large-scale compute is used to drive AI progress, the government can more effectively ensure AI benefits all of society.**

The remainder of this submission details two specific recommendations for how the government could do this, and the opportunity we believe the Future of Compute Review has in particular to address this key strategic priority for the UK.

Recommendation 1: Explore ways to increase compute access for academia

Most recent machine learning breakthroughs, and expected advances in this area, are reliant on large compute budgets. **Which groups have access to large-scale compute will therefore determine the incentives and interests that shape AI development**, and who is in a position to ensure its safety and trustworthiness.

In particular, there is a growing disparity between the computational resources available to AI researchers in academia and industry (Lohr 2019; Ahmed and Wahed 2020). **Figure A**, from a recent paper by Ganguli, et al, illustrates the trend of cutting-edge research skewing more and more toward industry.

Figure A. Industry's share of AI compute has grown at academia's expense

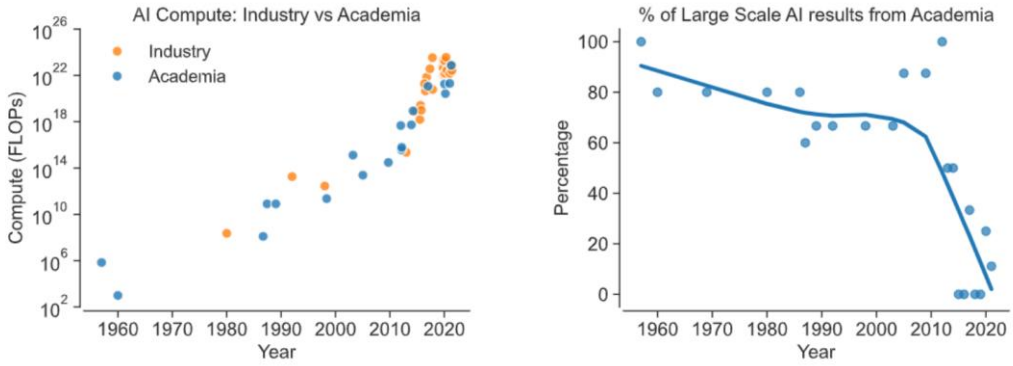

Figure 7 **(Left)** The amount of compute required by major AI projects over time is increasing exponentially for both academic (blue) and industrial (orange) projects. **(Right)** The proportion of computationally-intensive AI results from academia is steadily decreasing. (The blue curve represents a Lowess fit to the data.)

This means AI research is likely to be skewed towards short-term, commercial aims (for example, maximising click-through) rather than long-term benefits of AI for society as a whole. Increasing the computing power available to academics working in AI could have many benefits (Brundage et al 2020):

- **Increasing scrutiny of commercial models**: independent third-party stress-testing of models can be a valuable and effective way of catching possible failure modes before deployment. Academics are particularly well-placed to conduct such scrutiny, but often lack the computational resources or access to do so. With increased compute capacity, academics could replicate industry models making it much easier to test and scrutinise them.
- **Advancing beneficial AI applications** (such as medical research and diagnostics, AI for sustainable development goals) **and AI safety and security research**, for which free or subsidized compute resources could be provided specifically.
- **Supporting the UK's AI ecosystem**: by making it easier for more people to study and develop AI systems, we can increase the number of people with AI skills, which will have beneficial downstream effects on the economy.
- **Helping researchers train the next generation**: by reducing the number of leading researchers who are currently leaving academia for industry, where there is greater access to the compute resources to do world-leading research.

Other countries, including the US and Canada, are already looking into how cloud computing resources can be used to support national academic research. Canada's national advanced research computing (ARC) platform operates annual Resource Allocations Competitions (RAC) for eligible researchers. As of January 2022, their platform "serve[d] the needs of more than 19,000 users, including over 5,100 faculty based at Canadian institutions." Since 2016, the number of new applications received for each RAC has increased by an average of >10% annually. **Figure B.** below includes an illustrative graph and table created by the Digital Research Alliance of Canada.

Figure B. CPU request and allocation trends, Canada's RACs (2013-2022)

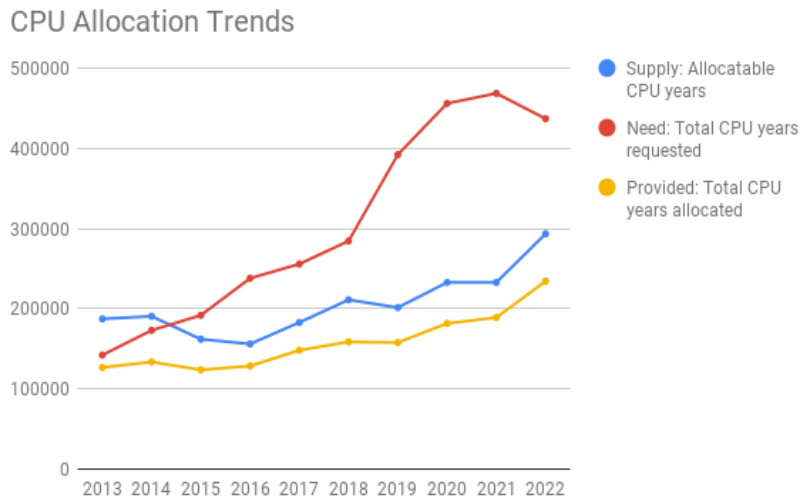

Table 3: Historical Compute Ask vs. Allocation

	Allocatable CPU CY	Total CYRequested	Total CYallocated	Difference	% of the demand awarded
2022	293,312	436,780	234,275	202,505	54%
2021	232,704	468,498	188,925	279,573	40%
2020	232,704	455,892	181,502	274,390	40%
2019	201,320	390,352	157,262	233,089	40%
2018	211,020	284,347	158,612	129,325	56%
2017	182,760	255,638	148,100	107,538	58%
2016	155,952	237,862	128,463	109,399	54%
2015	161,888	191,690	123,699	67,991	65%
2014	190,466	172,989	133,508	39,481	77%
2013	187,227	142,106	126,677	15,429	89%

As ARC's data shows, despite increased compute availability for Canadian researchers in recent years, demand growth has continually outpaced supply growth. Along with secular growth in compute-dependent research, previously latent demand for compute has also become visible as RAC-eligible researchers learned about and began applying for compute allocations.

Academic investigators rarely have access to the technical personnel and compute to fully realise their research agendas. This is true when it comes to vital AI safety research, and other areas of socially valuable research that are slowed by compute costs. A Canada RAC-style offering in the

UK would help stimulate further demand as new PIs "enter the market" and compete for access to compute, and develop socially valuable research that avails itself of compute resources.

In addition to providing greater compute access for academia, **the Review could also consider providing access to trained models via APIs**, making it easier for academics to learn from and scrutinise large-scale AI models without always needing access to large amounts of compute (Anderljung, Heim and Shevlane 2022). Using structured access, where developers facilitate controlled, arm's length interactions with their AI systems via an API, would make it possible for commercial models to receive greater scrutiny, potentially identifying security vulnerabilities, failure modes, or biases, without also increasing the risk that others misuse the models for harm (Shevlane 2022).

Along with APIs and awards of compute capacity, governments may unlock additional latent demand for compute by investing in enabling resources — e.g., technical staffers who can help researchers that are new to government compute resources to structure their projects. Researchers at Georgetown University's Center for Security and Emerging Technology noted in a 2022 report that training modern models "requires careful orchestration on a technical level," and that the demands of these models "requir[es] significant technical expertise to manage." Investing in technical support for research teams will enable a broader set of British investigators to compete for compute allocations, and to execute their projects post-award.

Recommendation 2: Consider monitoring and governing compute usage

In the coming years and decades, compute-driven AI progress is likely to transform many parts of society and the economy. In order to leverage the greatest possible benefits for society, the government must find ways to intervene in a timely and effective manner to prevent harm. However, given the rapid pace of progress and the fact that a majority of AI development occurs in the private sector, governments are generally only able to address harms after a capability has already been deployed in society and harms have often already occurred.

In order to build an effective, trusted AI governance regime, the UK government therefore needs both: (a) **better *information* about where potential harms might occur**, and (b) **more effective *tools* for intervening to prevent harm**.

Compute is a strategic leverage point which can help with both these challenges: since compute-intensive AI systems are particularly likely to precipitate unexpected harms, collecting data about compute usage can provide valuable information about where to focus anticipatory governance and risk assessment efforts. In the future, controlling how and where compute is used could also be an effective, practical lever for governments to shape AI progress.

As a crucial input into AI progress, compute is much more easily monitored and governed than other inputs such as data, algorithms, or talent. It is difficult to quantify how much additional progress is likely to result from adding a new person to a project, and difficult to control how talent is distributed across the AI ecosystem on anything more granular than the

state level. Both data and algorithms are very easy to copy, very easy to use for multiple purposes at once, and very highly distributed, making it extremely difficult to monitor how and where they are being used, or to prevent them being used for a specific purpose. By contrast, compute is much more easily quantifiable - there is a high correlation between the amount of compute being used by a project and the emergence of new capabilities - and much easier to monitor and control, since the supply and use of compute is highly centralized.

Figure C. below outlines how different potentially governable resources compare on different important dimensions (including excludability - how easy it is to control access to the research - and rivalry - whether using a resource in one context prevents it from being used in another, making it easier to track uses), illustrating how compute is the only resource that scores highly across all dimensions (Brundage and Sastry, in preparation - details available on request).

Figure C. Source: ***Brundage and Sastry, Computing Power and the Governance of Artificial Intelligence, in preparation***

Inputs					Training process	Output
	Training Data	Algorithms	Compute	Talent	→	Trained AI
Excludability	High	Mixed (cf. simultaneous invention)	High	High (though not nec. products of talent)		Mixed
Rivalry	Low	Low	High	High		Low
Quantifiability	Medium	Medium	High	Low		Mixed
Concentration of supply	Medium	Medium	High	Medium		Low
Concentration of use	Medium	Low	High	Medium		Mixed

Beginning to collect and act on information about compute usage could strongly support the government's aim of establishing a pro-innovation approach to regulating AI: by making it easier in future to systematically identify potentially high-risk capabilities ahead of time, the government can more effectively direct regulatory attention and risk-assessment to those capabilities, while leaving innovation in low-risk areas relatively unencumbered.

Monitoring and governing compute usage is also important if the UK government is to "accurately assess long term AI safety and risks" (National AI Strategy, p.60), especially given the potential emergence of much more radically transformative AI capabilities, including potentially Artificial General Intelligence (AGI) (Gruetzemacher and Whittlestone 2021). Such transformative capabilities are likely to be extremely compute intensive (Kaplan et al 2020;

Ganguli et al 2022), could emerge more rapidly than expected, and some actors may be motivated to keep them secret. For the UK government to be able to actively direct potentially transformative AI capabilities, it must build up capacity to identify such projects, and compute-usage monitoring is a very promising way to do so (Hwang 2018).

Below we outline a "tiered" approach to compute-indexed AI monitoring and governance, to illustrate what different levels of this could look like, and what would be needed to implement this in practice:

Tier 0: passive monitoring by academia and industry, general trends identified

- This is where we are now: academics and industry researchers in AI have the best information about compute use and capability development - governments only learn about AI capabilities after deployment, and even then only if deployed in public / published.
- Risk assessment is based on evidence of harm already having occurred or speculation about specific domains where harm might occur; no systematic way to identify potentially high-risk capabilities ahead of time.

Tier 0.5: active monitoring by government through existing information routes

- Government could collect data on broad compute trends, leveraging information provided from financial reporting, import duties, export controls, and information volunteered by AI companies and researchers in government-run foresight exercises.
- Even this fairly minimal level of monitoring could help with risk management and assessment in specific domains, or by detecting and stopping sanctioned actors from accessing compute (and creating an expectation of enforcement of such controls amongst compute providers).
- However, this would not allow governments to identify and tailor risk-assessment to specific systems, depends on voluntarily provided information, and would only provide rough proxies

Tier 1: active monitoring by government through mandatory AI developer and compute provider reporting

- Government could create reporting requirements for both users of compute (AI developers) and providers of compute (data center operators)
- The reporting threshold could be negotiated and revised to capture systems of interest (likely novel foundation models) while minimizing burden on private actors
- Reporting could leverage information already available at the technical level, in its native format (e.g. configuration files and logs of large training runs)
- Reported systems could, but don't have to, trigger further government inquiry, allowing anticipatory governance and per-system risk assessment
- Initially this would capture systems developed by UK-based actors or on UK-provided hardware, but wider global adoption (especially by the US), and information-sharing between governments, could capture a significant portion of novel AI capabilities

Tier 2: monitoring as in Tier 1 + mandatory risk assessment

- As domains, properties and thresholds for risks from AI systems become clearer, the government could mandate mandatory risk assessment for novel, compute-intensive AI systems.
- Risk assessment could be carried out by a 3rd party auditor.
- Risk assessment could also be carried out by trusted partners in academia, as long as they have sufficient expertise (which in turn likely depends on sustained access to sufficient compute, see recommendation 1).
- Risk assessment could leverage industry best-practices, and any standards or tools and benchmarks developed for such purposes (efforts to create such instruments are underway)

Tier 3: pre-development/pre-training risk assessment

- Theoretical research suggests that some future AI systems could pose unacceptable risks even during their development, though no such systems exist today.
- If there is growing consensus that such risks are near or imminent, or if empirical evidence of such risks emerge, the government could mandate pre-development risk assessment, and prevent actors from accessing the compute required for such development until risk assessment is cleared.
- To be in a position to enact such policies, the government would already need to be deeply familiar with the compute usage landscape, building on expertise, institutions and processes developed in lower tiers.

There are many important questions around the implementation of this proposal which we would be excited to see this review explore, and would be happy to provide further support around. These include:

- How much and what type of information on compute usage can be gathered from existing sources, and how this can be used to support the UK's current policy and regulatory goals;
- How to measure and compare compute usage in a standardised way across organisations;
- How to get buy-in from third parties for reporting requirements, and how to ensure or enable compliance;
- How the UK could set international norms and standards in this area;
- Which potential risks and harms might *not* be captured by focusing on compute-intensive AI development, and what additional measures might be needed to ensure these are not neglected.

References

Ahmed, N. and Wahed, M., 2020. The de-democratization of ai: Deep learning and the compute divide in artificial intelligence research. *arXiv preprint arXiv:2010.15581*.

Anderljung, M., Heim, L., and Shevlane, T. 2022. Compute Funds and Pre-trained Models. Available at: https://www.governance.ai/post/compute-funds-and-pre-trained-models

Angwin, J., Larson, J., Mattu, S., Kirchner, L. 2016. Machine Bias. ProPublica.

Amodei, D. and Hernandez, D. 2018. AI and Compute. Available at: https://openai.com/blog/ai-and-compute/

Bartlett, R., Morse, A., Stanton, R. and Wallace, N., 2022. Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics*, *143*(1), pp.30-56.

Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S., 2021, March. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?□. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623).

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E. and Brynjolfsson, E., 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B. and Anderson, H., 2018. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R. and Maharaj, T., 2020. Toward trustworthy AI development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*.

Clarke, S. and Whittlestone, J., 2022, July. A Survey of the Potential Long-term Impacts of AI: How AI Could Lead to Long-term Changes in Science, Cooperation, Power, Epistemics and Values. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 192-202).

Ganguli, D., Hernandez, D., Lovitt, L., Askell, A., Bai, Y., Chen, A., Conerly, T., Dassarma, N., Drain, D., Elhage, N. and El Showk, S., 2022, June. Predictability and surprise in large generative models. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1747-1764).

Gruetzemacher, R. and Whittlestone, J., 2022. The transformative potential of artificial intelligence. *Futures*, *135*, p.102884.

Hwang, T., 2018. Computational power and the social impact of artificial intelligence. *arXiv preprint arXiv:1803.08971*.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. and Amodei, D., 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Lohr, S., 2019. At tech's leading edge, worry about a concentration of power. *The New York Times*, *26*, p.2019.

Nelson, G. 2015. Google, Delphi disclose crashes in self-driving cars. Automotive News.

Seger, E., Avin, S., Pearson, G., Briers, M., Heigeartaigh, S.Ó., Bacon, H., Ajder, H., Alderson, C., Anderson, F., Baddeley, J. and Bakker, C., 2020. Tackling threats to informed decision-making in democratic societies: Promoting epistemic security in a technologically-advanced world.

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M. and Villalobos, P., 2022. Compute trends across three eras of machine learning. *arXiv preprint arXiv:2202.05924*.

Shevlane, T., 2022. Structured Access: An Emerging Paradigm for Safe AI Deployment. The Oxford Handbook of AI Governance, Justin Bullock (ed.) et al.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A. and Kenton, Z., 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Whittlestone, J. and Clarke, S., 2022. AI Challenges for Society and Ethics. The Oxford Handbook of AI Governance, Justin Bullock (ed.) et al.

Wiens, J., Price, W.N. and Sjoding, M.W., 2020. Diagnosing bias in data-driven algorithms for healthcare. *Nature medicine*, *26*(1), pp.25-26.