T20 Policy Brief

G20 BRASIL 2024

T20 Brasil 2024 Let's rethink the world

Task Force 05

INCLUSIVE DIGITAL TRANSFORMATION

Governing Computational Infrastructure for Strong and Just AI Economies

Jai Vipra, Researcher, IT for Change (India)

Carla Rodrigues, Digital Platforms and Markets Coordinator, Data Privacy Brasil (Brazil)

Bruno Bioni, Founder and Director, Data Privacy Brasil (Brazil)

Rafael Zanatta, Executive Director, Data Privacy Brasil (Brazil)

Alison Gillwald, Executive Director, Research ICT Africa (South Africa)

Boxi Wu, Graduate Research Student, Oxford Internet Institute (United Kingdom)

Tanvi Lall, Director, Strategy, people+ai (India)

Swaroop Rajagopalan, Volunteer, people+ai (India)

TF05

Abstract

The development of AI systems today is constrained by the availability of computational power. AI-relevant computational capacity is supplied by extraordinarily concentrated markets. Large AI models use an ever-increasing amount of computational power and differentiate themselves with the use of the highest number of the most advanced chips. The use of these chips is further constrained through preferential access via vertical integration, and a concentrated cloud market helmed by Big Tech.

As the economic importance of AI continues to grow rapidly, access to computational power is being positioned by industry as potentially mediating production and exchange relationships, and by extension socio-economic well-being and political decision-making. The compute divide between academia and industry is reflected and amplified between the global south and north, and between the public and private sectors. Advanced computational resources now also feature prominently in geopolitical faultlines such as export controls and 'de-risking'. Computational power has emerged as a constraint for the development of digital public infrastructure that is independent of Big Tech. As more governments seek to build and direct a public digital innovation ecosystem, the question of access to computational power for public welfare becomes significant.

This policy brief recommends that the G20 come to an agreement on governing computational resources with a commitment to open compute paradigms, no remote hardware control mechanisms without consensus, and a serious assessment of the environmental impact of large computational infrastructure. Such an agreement is aimed at ensuring that every country is able to build strong and resilient domestic AI economies in a manner that promotes economic justice.

Keywords: compute, AI, competition, antitrust, compute governance, democratic AI, digital sovereignty

Diagnosis of the Issue

AI development today is dominated by the private sector. This domination is largely due to, and evident in, the accumulation of computational resources, including AI chips, supercomputers and the data centres that house them, by the private sector. The largest AI model trained in academia uses less than 1 per cent of the compute used by the largest AI model trained in industry (Besiroglu et al. 2024). Entire nation-states' public investments in AI compute are dwarfed by start-up investments in AI compute.

The cost of compute is extraordinarily high. So far, Gemini Ultra is the most expensive AI model, having cost about USD 630 million to train (Epoch 2023). The cost of compute is high in part due to the materials and expertise required, but also in large part due to the concentration in the semiconductor supply chain (Khan, Peterson, and Mann 2021). Only a handful of firms can design and produce the compute that is required to train large AI models.

In turn, compute drives extraordinary concentration downstream in AI markets as well. All the notable providers of large AI models today are existing technology giants. The cost of compute has driven mergers, acquisitions, and the overall leadership of Big Tech in AI markets. Any AI startup would have to depend on cloud service providers such as Amazon Web Services, Google Cloud, Microsoft Azure, Alibaba, Tencent and Huawei, or directly and indirectly on chip giants like Nvidia, Intel, and TSMC. We are seeing increasing vertical integration across chip design, production, cloud services, AI model development, platforms, and downstream AI-enabled products and services. The market concentration of digital platforms is seamlessly morphing into the market concentration of AI.

Unsurprisingly, we now witness a global skew in AI capabilities. The most advanced AI models are trained in the United States and to some extent in China. Promising competitors in the UK and the EU have been either acquired by US Big Tech corporations or have seen large investments from the latter. The Global South lags far behind in the development of AI models, and its role remains relegated to the provision of inputs kept at lowered values, like data annotation (Muldoon and Wu 2023; Png 2022). Much of the talent and materials required to develop advanced AI are drawn from the Global South, but this development takes place in the Global North, particularly within Big Tech (Thornhill 2024).

For Global South countries it is especially true that large public investments in AI compute take away from investments in healthcare, education, and other economic activity. Countries are caught in a bind – if they do not invest in AI compute, they risk being left far behind; if they do, they risk not succeeding, or the technology not providing adequate returns, or neglecting more urgent investment needs. Global South countries are also more susceptible to dependence on the infrastructure of Global North countries, which further exacerbates the compute divide and reinforces imperial relations (Kwet 2019).

The result of concentrated AI compute markets is unilateral decision-making in AI with universal ramifications. As things stand, the trajectory of AI development is determined by only a few people and their specific incentives, there are single points of failure at various stages, and a few countries' actions might determine the global future for decades to come.

Even as the decision-making becomes unilateral, the universal ramifications of AI are felt by people in their daily lives. Due to global value chains, AI used in one jurisdiction leads to extensive job losses in another place; the environmental impact of data centres

cannot be contained geographically; and harmful use cases of AI are rapidly imitated at a global scale.

Recommendations

An agreement on compute governance at the G20 can reduce the unilateralism of decision-making and the universality of ramifications of AI development today. G20 countries must take the lead for a just global future.

A G20 agreement on compute governance can set the norms for the global governance of AI compute on just and equitable lines. It will serve as a step towards overcoming geopolitical rivalries to build consensus on an issue of shared, global importance. An international agreement on AI compute – broader than the G20 itself – is vital because of the positive and negative externalities inherent in AI compute development and distribution.

We recommend that this agreement have the following commitments:

1. **Cooperating to develop, and investing in, open compute paradigms:** At least part of the high cost of compute development arises from the concentration in compute markets. Instead of redirecting valuable social resources towards rewarding this concentration, G20 governments must work together to develop alternatives through competition policy, regulation, and material support for open compute paradigms. Policies must incentivise the unbundling of compute software and compute hardware to promote competitive markets (Vipra and Myers West 2023)

Governments must commit to investing in open source compute software, as well as experiments in building digital public infrastructure for AI compute (people+ai 2024). Governments must encourage the development of open protocols for cloud compute.

Governments must also explore cooperative regional planning for AI compute through such an agreement, such as through shared infrastructure and decentralised compute facilities. It is not feasible for every country to make large investments in AI compute, and regional cooperative can go a long way in making such investments financially prudent.

Such a commitment would go a long way towards addressing both the extraordinary concentration in the compute supply chain, and the geographical skew in compute production and provision.

2. **Prohibiting the use of hardware control mechanisms without consensus:** As a regulatory mechanism, a few politicians, policy professionals and special interest groups have suggested and/or examined the use of on-chip hardware mechanisms that allow for remote monitoring and shutdown of compute clusters (Schumer et al. 2024; Arne, Fist, and Withers 2024; Reinsch and Benson 2021; Muehlhauser 2023; Sastry et al. 2024). We believe that such measures have more harms than benefits for three reasons:

a. They impinge upon the sovereignty of member states and other states as they allow for undue surveillance by foreign governments under the guise of national security. Such measures would allow governments to remotely disable other countries' AI systems, a capability that would be unacceptable to any country and that would erode trust in the international governance of compute.

b. Such proposed measures encourage every country or region to develop its own delinked compute system. Not only is this impractical for most countries, it encourages

the reckless and unnecessarily expensive development of compute, diverting resources, as mentioned above, from other important social goals.

c. Such proposed measures intensify state surveillance and control over individual computing activities. Hardware freedom is an important principle even if it is not absolute; giving governments sweeping powers over the very hardware of programming sets a poor precedent for a digital future.

We recognise that countries may mutually determine to institute such on-chip remote control mechanisms for narrow and high-risk activities such as the development of lethal autonomous weapons systems. In such narrowly defined agreements, remote verification measures might increase mutual trust.

3. **Addressing the environmental impact of large AI models:** It is becoming clearer that the training and development of large-scale AI models requires large computational and storage clusters which have vast environmental costs (OECD 2022). A single advanced AI chip can consume more energy than the average US household (Shilov 2023). Conservative estimates assess that data centres will consume 4.5% of global energy by 2030 (Patel, Nishball, and Ontiveros 2024). The water consumption of AI data centres has also received much journalistic, scholarly, and activist attention (Hogan 2015; Valdivia 2022; Hao 2024). G20 countries should therefore:

a. Commission a global study on the environmental impact of AI across its supply chain, with a special focus on compute infrastructure, i.e., chip production and data centres;

b. Commit to developing a system of monetary compensation for communities and countries affected by the establishment of data centres; and

c. Evolve limits on energy and water availability for very large data centres, including through assessments of social costs and social benefits of such data centres.

d. Explore the viability of smaller AI models for specific use cases.

Scenario of Outcomes

1. **Stifling of innovation** – It is possible that an international compute governance agreement with its controls and regulations, may discourage AI innovation. However, consider the following mitigating factors:

a. Through antitrust and other related measures, the concentration in the semiconductor supply chain can be broken, promoting competition. Competition can spur innovation, increase the supply of compute, and tie this supply to specific demand rather than to speculative ventures. This may also reorient AI development away from destructive geostrategic ventures towards socially beneficial AI innovation.

b. While larger models (i.e., those using more compute) have so far demonstrated greater capabilities than smaller models, the proportion of capability improvement in relation to model size increases is unclear. There are already strong doubts in the industry about size leading to greater capabilities at this point. This is an ideal juncture in the course of AI development to use regulation to re-assess the wisdom of investing very large amounts of money and natural resources into large models.

c. The current paradigm ensures that AI innovation takes place only within Big Tech, and if it does take place outside, it is quickly absorbed by Big Tech (Lehdonvirta 2024).

The proposed agreement on compute will promote innovation outside Big Tech even if it reduces incentives for some narrow innovation within Big Tech.

2. **Easy compute availability leads to AI misuse** – It is possible that a competitive and open compute ecosystem leads to the proliferation of harmful AI systems, including in critical contexts like war and surveillance. In the first instance, we recommend that an international agreement on autonomous weapons systems be arrived at on priority. Secondly, we note that concentrated and closed compute supply chains do not prevent the harmful use of AI, but rather increase the cost of beneficial uses and alternative directions for the technical development of AI.

3. **A more democratic, open and just AI ecosystem is developed** – If a G20 agreement on compute is arrived at based on the contours recommended above, we expect that a new AI ecosystem will develop. This ecosystem will be more democratic if only because it takes into consideration the goals of various governments, even if it is not directly governed by the people of the world. It will also be more open, allowing for different forms of innovation not constrained by the narrow motives of a handful of firms. Additionally, such an ecosystem will be more just because it considers both people and planet in its design.

References

Arne, Onni, Tim Fist, and Caleb Withers. 2024. 'Secure, Governable Chips'. Center for a New American Security. https://www.cnas.org/publications/reports/secure-governable-chips.

Besiroglu, Tamay, Sage Andrus Bergerson, Amelia Michael, Lennart Heim, Xueyun Luo, and Neil Thompson. 2024. 'The Compute Divide in Machine Learning: A Threat to Academic Contribution and Scrutiny?' arXiv https://doi.org/10.48550/arXiv.2401.02452.

Epoch. 2023. 'Machine Learning Trends'. Epoch. 11 April 2023. https://epochai.org/trends.

Hao, Karen. 2024. 'AI Is Taking Water From the Desert'. *The Atlantic*, 1 March 2024. https://www.theatlantic.com/technology/archive/2024/03/ai-water-climate-microsoft/677602/.

Hogan, Mél. 2015. 'Data Flows and Water Woes: The Utah Data Center'. *Big Data & Society* 2 (2): 2053951715592429. https://doi.org/10.1177/2053951715592429.

Khan, Saif M., Dahlia Peterson, and Alexander Mann. 2021. 'The Semiconductor Supply Chain'. Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/the-semiconductor-supply-chain/ .

Kwet, Michael. 2019. 'Digital Colonialism: US Empire and the New Imperialism in the Global South'. *Race & Class* 60 (4): 3–26. https://doi.org/10.1177/0306396818823172 .

Lehdonvirta, Vili. 2024. *Cloud Empires: How Digital Platforms Are Overtaking the State and How We Can Regain Control*. Cambridge, Massachusetts London, England: The MIT Press.

Muehlhauser, Luke. 2023. '12 Tentative Ideas for US AI Policy | Open Philanthropy'. *Open Philanthropy* (blog). 17 April 2023. https://www.openphilanthropy.org/research/12-tentative-ideas-for-us-ai-policy/.

Muldoon, James, and Boxi A. Wu. 2023. 'Artificial Intelligence in the Colonial Matrix of Power'. *Philosophy & Technology* 36 (4): 80. https://doi.org/10.1007/s13347-023-00687-8.

OECD. 2022. 'Measuring the Environmental Impacts of Artificial Intelligence Compute and Applications: The AI Footprint'. OECD Digital Economy Papers 341. Vol. 341. OECD Digital Economy Papers. https://doi.org/10.1787/7babf571-en.

Patel, Dylan, Daniel Nishball, and Jeremie Eliahou Ontiveros. 2024. 'AI Datacenter Energy Dilemma - Race for AI Datacenter Space'. 13 March 2024. https://www.semianalysis.com/p/ai-datacenter-energy-dilemma-race.

people+ai. 2024. 'GitHub - PeoplePlusAI/Open-Cloud-Compute-OCC'. GitHub. https://github.com/PeoplePlusAI/Open-Cloud-Compute-OCC.

Png, Marie-Therese. 2022. 'At the Tensions of South and North: Critical Roles of Global South Stakeholders in AI Governance'. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1434–45. FAccT '22. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3531146.3533200.

Reinsch, William Alan, and Emily Benson. 2021. 'Digitizing Export Controls: A Trade Compliance Technology Stack?' Center for Strategic and International Studies. https://www.csis.org/analysis/digitizing-export-controls-trade-compliance-technology-stack.

Sastry, Girish, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, Julian Hazell, Cullen O'Keefe, et al. 2024. 'Computing Power and the Governance of Artificial Intelligence'. arXiv. https://doi.org/10.48550/arXiv.2402.08797.

Schumer, Chuck, Mike Rounds, Martin Heinrich, and Todd Young. 2024. 'Driving US Innovation in Artificial Intelligence'. The Bipartisan Senate AI Working Group, United States Senate. https://www.schumer.senate.gov/imo/media/doc/Roadmap_Electronic1.32pm.pdf.

Shilov, Anton. 2023. 'Nvidia's H100 GPUs Will Consume More Power than Some Countries — Each GPU Consumes 700W of Power, 3.5 Million Are Expected to Be Sold in the Coming Year'. *Tom's Hardware* (blog). 26 December 2023. https://www.tomshardware.com/tech-industry/nvidias-h100-gpus-will-consume-more-power-than-some-countries-each-gpu-consumes-700w-of-power-35-million-are-expected-to-be-sold-in-the-coming-year.

Thornhill, John. 2024. 'How Big Tech Is Winning the AI Talent War'. *Financial Times*, 23 March 2024. https://www.ft.com/content/2892bac2-d848-49ea-b983-bc649a8c0529.

Valdivia, Ana. 2022. 'Silicon Valley and the Environmental Costs of AI'. *Political Economy Research Centre* (blog). 5 December 2022. https://www.perc.org.uk/project_posts/silicon-valley-and-the-environmental-costs-of-ai/.

Vipra, Jai, and Sarah Myers West. 2023. 'Computational Power and AI'. AI Now Institute. https://ainowinstitute.org/publication/policy/compute-and-ai.

G20 BRASIL 2024
BUILDING A JUST WORLD AND A SUSTAINABLE PLANET

T20 Brasil 2024
Let's rethink the world

Let's rethink the world

CEBRI
Rethink Tank

FUNDAÇÃO ALEXANDRE DE GUSMÃO

ipea
Instituto de Pesquisa Econômica Aplicada