OECD AI Compute Taskforce

Process overview note by Keith Strier, Chair of the OECD AI Compute Taskforce.

AI is set to affect a growing number of economic sectors creating new opportunities, improving productivity and changing the labour-capital mix. Understanding countries' capacity and readiness to embrace this fast-evolving transition is essential. One of the conditions for the deployment of AI applications is the availability of relevant infrastructure enabling computation at scale.

Figure 1. AI enablers

Source: OECD

For this reason, the OECD Observatory for AI Policy is expanding its Network of Experts on AI (ONE AI) to launch a forthcoming AI Compute Taskforce ("Taskforce"). The aim of the Taskforce will be to provide consistent data and indicators on countries' respective endowment and investment in the *computational infrastructure* or *computing capacity* enabling AI systems' development (thereafter, "compute"). In doing so, it will attempt to help countries answer three questions fundamental to implementing national AI strategies:

- How much AI compute does the country have?
- How does it compare to other countries?
- Is current AI compute capacity sufficient to support national AI development objectives?

Understanding national capacity in computational infrastructure to calibrate and deploy AI-related policies is essential. Yet governments seizing investment plans in AI-enabling compute infrastructure tend to have relatively little information, mainly due to the absence of widely accepted definitions, standards, and benchmarks on AI compute.

There are several reasons for this lack of common measurement tools. Firstly, AI policy efforts have so far been focused on data and algorithms, partly overlooking the underlying compute infrastructure that enables AI develpoment. AI compute is also a complex field, involving a specialised stack of hardware and software growing particularly rapidly. Adding further challenge is the fact that AI compute can be accessed in various ways: as a service through a private or public cloud, using a purpose-built data science workstation and even "at the edge" on mobile devices (see below). Compute is also applied in many different fields and domains to solve complex problems such as particle physics, vaccine development, public health monitoring, financial trading, natural language modeling and weather forecasting. As a result, any assessment of total AI compute by country requires analysts to consider a variety of sectors, industries, platforms, configurations and delivery methods.

To bridge the gap in data availability, the Taskforce intends to produce a common OECD Framework for measuring and benchmarking domestic AI compute capacity, with the aim to inform national assessments of AI readiness and national AI investment priorities. The Framework will be built around three main elements:

1. A user-friendly framework for measuring domestic AI compute capacity

2. A benchmarking tool to help countries compare domestic AI compute capacity

3. A model to gauge the sufficiency of domestic AI compute capacity

Developing a user-friendly framework for measuring domestic AI compute capacity

Policy makers working on AI-related issues will inevitably need to make informed decisions about a technical topic without having necessarily received formal training in data science or related fields. As a result, and similarly to the OECD Framework for the Classification of AI Systems, this Framework aims to be a user-friendly tool for policy makers facing such decisions on the topic of AI compute infrastructure. It does not aim to be exhaustive.

Benchmarking domestic AI compute capacity

The Framework will bring actionable insights to policymakers if they can also compare data across countries. For this reason, it is critical that the Framework's development be guided by feasibility concerns on the types of data may be practically sourced and consistently tracked across countries.

A visual benchmarking tool based on comparable international data will help policymakers better understand their national competitiveness in the context of the global AI landscape as well as what is needed to ensure their digital sovereignty.

Assessing the sufficiency of domestic AI compute capacity

The Taskforce's final objective will to enable decision makers to link AI compute capacity to specific policy goals. As such, the Taskforce aspires to help countries understand whether their domestic AI compute capacity is sufficient to advance AI and/or economic policy objectives, such as:

- Implementing national AI plans
- Supporting economic growth goals
- Competing in a global digital economy
- Advancing the national science agenda
- Detecting critical infrastructure cyber-threats
- Accelerating responses to natural disasters
- Enabling corporate innovation to support legacy industries
- Attracting start-ups to fuel market innovation

To advance towards these three overarching objectives, the Taskforce will engage ONE AI experts and stakeholders to try to answer key theoretical and practical questions underpinning measurement of AI compute. Some of these questions are listed below:

Key methodological and conceptual questions

WHAT IS AI COMPUTE?

AI compute is a specialized stack of hardware and software involving processors or chips, servers, storage, software, and networking, all designed to support AI-specific workloads and applications. It thus covers a range of different technologies from AI chips to data servers to cloud computing.

AI development has two phases: training and inference. In the first phase, an AI model is "trained" on data. In the second phase, a trained AI model is deployed and then "infers" (i.e., makes decisions or take actions) in the field based on new data. Training and inference phases run on AI compute infrastructure in a private data center or in the public cloud. Everything outside the data center (or cloud) is typically considered the "edge." This includes a vast constellation of smart, connected sensors, devices, and machines (phones, cameras, self-driving trucks, satellites, delivery robots, drones, etc.). This constellation is also referred to as the "Internet of Things."

HOW CAN ONE DELINEATE COUNTRIES' RESPECTIVE COMPUTE CAPACITY?

An important definitional issue to address will be how to delineate different countries' AI compute stocks. There is currently no consensus, on what qualifies as "domestic" AI compute. The Taskforce will need to consider if AI compute can be classified as domestic if it is 1) owned and operated by a non-domestic private sector actor and/or 2) physically located in another country.

Aggregating the performance of individual AI systems within a country is one way to calculate that country's "AI compute capacity" but there are limitations to this approach. Commonly used benchmarks are narrowly formulated to define performance under very strict conditions and may not be applicable to all AI systems within a country.

Another approach is to count the number of discreet AI systems and group them by "class" of performance, such as leadership-class AI systems and center-class AI systems. This approach may provide less-specific results but has the benefit of being more user-friendly.

Given the paucity of good data on AI systems by country, the Taskforce may also consider a composite index approach that combines multiple inputs to generate a score. The Taskforce is also considering novel ways to leverage existing benchmarks, such as MLPerf, promoting new kinds of submissions to generate a new data set to better inform policymaking in the years ahead.

SHOULD AI COMPUTE CAPACITY IN THE PUBLIC CLOUD BE MEASURED?

Wikipedia defines cloud computing as the "on-demand availability of computer system resources". As such, a cloud is a service built around a data center made available to many users, typically through the Internet, which can be intended for private use by a defined set of users (employees, etc.) or for public use by commercial customers. There are many different types of "clouds" based on who operates them and for what purpose, including:

- Public Cloud is a commercial infrastructure for paying customers
- Private Cloud is a closed infrastructure for affiliated Business Units, employees, and vendors
- Government Cloud is a closed infrastructure for gov't agencies, staff, and contractors
- Research Cloud is a closed infrastructure for affiliated R&D labs and researchers
- Sovereign Cloud is a state-sponsored/subsidized service for citizens

It is increasingly common for governments to pursue a hybrid national AI compute strategy that combines 1) investments in domestic cloud and AI compute infrastructure to maintain infrastructure sovereignty with 2) partnerships with public cloud service providers to capitalize on private sector innovation and broaden access to AI compute within their borders.

Measuring public cloud AI compute capacity by country may be challenging, however, given the borderless nature of the service provided. That said, it may still be helpful to policy makers for the OECD to benchmark the extent to which member countries leverage the public cloud as a strategic component of their national AI roadmap. This may be achieved by collecting data on 1) national AI roadmaps that explicitly incorporate the public cloud, 2) announced partnerships with public cloud service providers and 3) government AI workloads targeted for the public cloud.

SHOULD AI COMPUTE CAPACITY OUTSIDE THE DATA CENTER BE MEASURED?

Over the next few years, it is predicted that 80% of AI compute in the world will in fact move outside data centers (and the cloud) to "the edge". Given the scale, complexity, and impracticality of measuring AI compute across trillions of smart, connected devices, the Taskforce will initially focus on data center-based AI compute capacity.

An interesting follow-on task may be to assess the potential for aggregating certain classes of edge AI compute as an input to national AI compute capacity. There is a precedent for this notion with personal computers. For example, one of the largest supercomputers in the world today is the virtualized AI system comprised of hundreds of thousands of personal gaming computers that have been connected through the internnet to the Folding@Home initiative (https://foldingathome.org/).

Conclusion

AI is a general-purpose technology impacting every facet of the global economy, prompting a worldwide imperative for governments to formulate and publish comprehensive national AI strategies. Those same governments are already allocating budgets and investing public funds – in the billions to hundreds of billions per country - to support the implementation of AI programs in pursuit of what many studies estimate to be trillions in net new economic value added[1]. As such, the successful implementation of a national AI strategy may become one of the defining factors in a country's ability to innovate, support productivity and long-term growth.

Given this context, the lack of widely accepted definitions, standards, and data sets on which to develop evidence-based public policy for AI compute calls for renewed effort and cooperation. The OECD Taskforce thus embraces the opportunity to close this gap with a view to empower OECD countries to make more informed, data-driven policy decisions.

[1] https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf