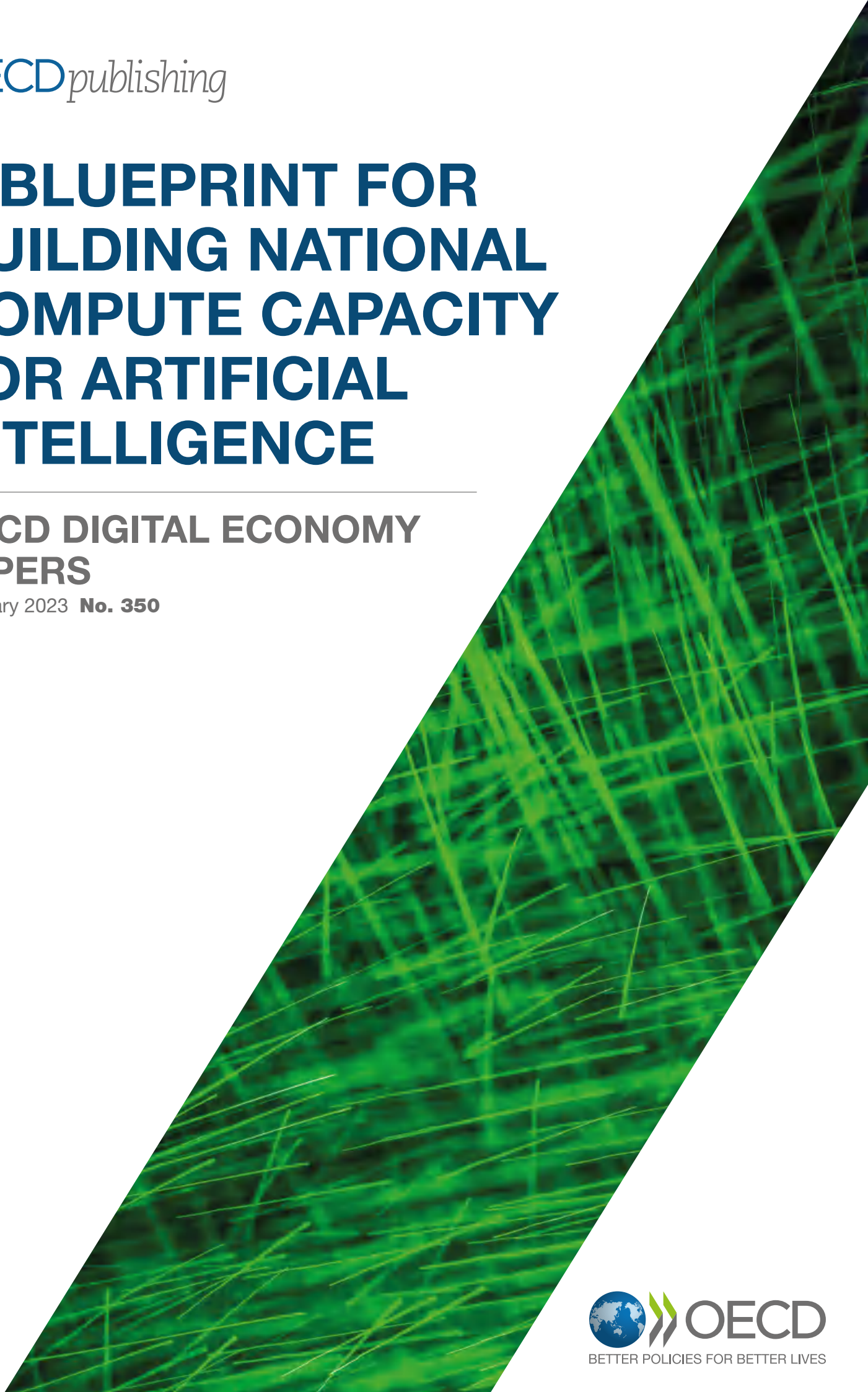
OECD publishing

A BLUEPRINT FOR BUILDING NATIONAL COMPUTE CAPACITY FOR ARTIFICIAL INTELLIGENCE

OECD DIGITAL ECONOMY PAPERS
February 2023 **No. 350**



OECD
BETTER POLICIES FOR BETTER LIVES

This paper was prepared for publication by the OECD Secretariat in consultation with the OECD.AI Expert Group on AI Compute and Climate. The paper was approved and declassified by the Committee on Digital Economy Policy (CDEP) on 23/12/2022.

Note to Delegations:

This document is also available on O.N.E under the reference code:

DSTI/CDEP/AIGO(2022)2/FINAL

Abstract

Artificial intelligence (AI) is transforming economies and promising new opportunities for productivity, growth, and resilience. Countries are responding with national AI strategies to capitalise on these transformations. However, no country today has sufficient data on, or a targeted plan for, national AI compute capacity. This policy blind-spot may jeopardise domestic economic goals. This report provides the first blueprint for policy makers to help assess and plan for the national AI compute capacity needed to enable productivity gains and capture AI's full economic potential. It provides guidance for policy makers on how to develop a national AI compute plan along three dimensions: capacity (availability and use), effectiveness (people, policy, innovation, access), and resilience (security, sovereignty, sustainability). The report also defines AI compute, takes stock of indicators, datasets, and proxies for measuring national AI compute capacity, and identifies obstacles to measuring and benchmarking national AI compute capacity across countries.

Abrégé

L'intelligence artificielle (IA) transforme les économies et les sociétés et ouvre la voie à de nouvelles perspectives en termes de productivité, de croissance et de résilience. Dans une volonté de capitaliser sur ces transformations, les pays définissent des stratégies nationales en matière d'IA. En revanche, ils omettent souvent de déterminer s'ils disposent d'une capacité de calcul pour l'IA suffisante pour atteindre leurs objectifs nationaux, concrétiser les gains de productivité et exploiter le plein potentiel économique de l'IA. Ce rapport dessine un cadre destiné à aider les décideurs à définir des plans nationaux en matière de capacité de calcul pour l'IA qui soient cohérents avec les stratégies et besoins de leur pays dans le domaine de l'IA. Il définit la notion de capacité de calcul pour l'IA et recense les indicateurs, les ensembles de données et les variables de substitution permettant de mesurer la capacité de calcul nationale. Il indique ensuite aux décideurs comment évaluer les besoins technologiques et définir lesdits plans nationaux en prenant en considération la capacité de calcul (disponibilité et utilisation), l'efficacité (ressources humaines, politique, innovation, accès) et la résilience (sécurité, souveraineté, durabilité). Le rapport fait également le point sur les obstacles à la mesure et l'analyse comparative des capacités de calcul pour l'IA entre les pays.

Executive summary

Artificial intelligence (AI) is transforming economies and promising new opportunities for productivity, growth, and resilience. Embracing AI-enabled transformation depends on the availability of infrastructure and software to train and use AI models at scale. Ensuring countries have sufficient such "AI compute capacity" to meet their needs is critical to capturing AI's full economic potential.

Many countries have developed national AI strategies without fully assessing whether they have sufficient domestic AI compute infrastructure and software to realise their goals. Other AI enablers, like data, algorithms, and skills, receive significant attention in policy circles, but the hardware, software, and related infrastructure that make AI advances possible have received comparatively less attention. Today, standardised measures of national AI compute capacity remain a policy gap. Such measures would give OECD and partner economies a greater understanding of AI compute and its relationship to the diffusion of AI, improve the implementation of AI strategies, and inform future policy and investments.

The demand for AI compute has grown dramatically for machine learning systems, especially deep-learning and neural networks. According to research, the computational capabilities required to train modern machine learning systems, measured in number of mathematical operations (i.e., floating-point operations per second, or FLOPS), has multiplied by hundreds of thousands of times since 2012[1] (OpenAI, 2018[1]; Sevilla et al., 2022[2]), despite algorithmic and software improvements that reduce computing power needs. The increasing compute needs of AI systems create more demand for specialised AI software, hardware, and related infrastructure, along with the skilled workforce necessary to utilise them efficiently and effectively.

As governments invest in developing cutting-edge AI, compute divides can emerge or deepen. An imbalance of such compute resources risks reinforcing socioeconomic divides, creating further differences in competitive advantage and productivity gains. Over the past decade, private sector led initiatives within countries have increasingly benefitted from state-of-the-art AI compute resources, particularly from commercial cloud service providers, compared to public research institutes and academia. The OECD.AI Expert Group on AI Compute and Climate advances collective understanding and measurement of AI compute to shed light on AI compute divides between countries and within national AI ecosystems.

This report offers a blueprint for policy makers to develop national AI compute plans aligned with national AI strategies and domestic needs. It takes stock of existing and proposed indicators, datasets, and proxies for measuring national AI compute capacity. Policy makers can assess technology needs and develop national AI compute plans by considering compute's capacity (availability and use), effectiveness (people, policy, innovation, access), and resilience (security, sovereignty, sustainability).

Findings and measurement gaps are identified to inform future work in developing AI-specific metrics to quantify and benchmark AI compute capacity across countries. They include: national AI policy initiatives need to take AI compute capacity into account; national and regional data collection and measurement standards need to expand; policy makers need insights into the compute demands of AI systems; AI-specific measurements should be differentiated from general-purpose compute; workers need access to AI compute related skills and training for effective AI compute use; and AI compute supply chains and inputs need to be mapped and analysed.

Résumé

L'intelligence artificielle (IA) transforme les économies et les sociétés et ouvre la voie à de nouvelles perspectives en termes de productivité, de croissance et de résilience. La capacité d'un pays à amorcer une transformation fondée sur l'IA dépend de la disponibilité des infrastructures et des logiciels nécessaires pour entraîner et utiliser les modèles d'IA à grande échelle. Pour exploiter le plein potentiel économique de l'IA, il est impératif que les pays disposent de capacités de calcul à la hauteur de leurs besoins.

De nombreux pays ont défini des stratégies nationales en matière d'IA sans véritablement déterminer s'ils possèdent à l'échelle nationale des infrastructures et logiciels suffisants pour atteindre leurs objectifs. Si les pouvoirs publics s'intéressent de près à un certain nombre de facteurs qui sous-tendent l'IA, comme les données, les algorithmes et les compétences, ils prêtent une attention relativement moindre au matériel, aux logiciels et à l'infrastructure connexe indispensables aux progrès de l'IA. On ne dispose pas à l'heure actuelle de mesures normalisées des capacités nationales de calcul pour l'IA. De telles mesures aideraient les pays de l'OCDE et les économies partenaires à mieux appréhender la capacité de calcul et sa corrélation avec la diffusion de l'IA, à améliorer la mise en œuvre des stratégies en matière d'IA et à guider les politiques et les investissements futurs.

On assiste à une explosion des besoins en capacité de calcul pour les systèmes d'apprentissage automatique, en particulier les réseaux neuronaux et l'apprentissage profond. Des travaux de recherche ont montré que les capacités de calcul requises pour entraîner des systèmes d'apprentissage automatique modernes, mesurées en nombre d'opérations mathématiques (c'est-à-dire en nombre d'opérations en virgule flottante par seconde, ou flops) sont des centaines de milliers de fois supérieures à celles de 2012[1] (OpenAI, 2018[1]; Sevilla et al., 2022[2]), bien que les progrès des algorithmes et des logiciels aient permis de réduire les besoins en puissance de calcul. Les besoins croissants en capacité de calcul des systèmes d'IA font progresser la demande de logiciels spécialisés dans l'IA, de matériel et d'infrastructures connexes, ainsi que d'une main-d'œuvre qualifiée capable de les utiliser de manière efficiente et efficace.

À mesure que les pouvoirs publics investissent dans la conception de systèmes d'IA de pointe, des disparités de capacités de calcul peuvent apparaître ou s'aggraver. Les déséquilibres liés à ces ressources risquent de creuser les fractures socio-économiques existantes et, par ricochet, d'accentuer les écarts de compétitivité et de productivité. Au cours des dix dernières années, les initiatives menées dans les pays par des acteurs du secteur privé ont davantage bénéficié de ressources de calcul ultramodernes, fournies notamment par des prestataires de services infonuagiques commerciaux, que les établissements publics de recherche et les universités. Le Groupe d'experts OECD.AI sur la capacité de calcul pour l'IA et le climat s'attache à faire progresser la compréhension commune et la mesure de la capacité de calcul pour mettre en évidence les écarts entre les pays et au sein des écosystèmes d'IA nationaux.

Le présent rapport propose un cadre visant à aider les décideurs à définir des plans nationaux en matière de capacité de calcul pour l'IA qui soient cohérents avec les stratégies et besoins de leur pays dans le domaine de l'IA. Il recense les indicateurs, les ensembles de données et les variables de substitution existants et envisagés pour mesurer la capacité de calcul nationale pour l'IA. Les décideurs peuvent évaluer les besoins technologiques et définir lesdits plans nationaux en prenant en considération la capacité de calcul (disponibilité et utilisation), l'efficacité (ressources humaines, politique, innovation, accès) et la résilience (sécurité, souveraineté, durabilité).

Les lacunes en termes de mesure et les conclusions qui ressortent du rapport pourront nourrir de prochains travaux et aider à élaborer des indicateurs permettant de quantifier et de comparer la capacité de calcul pour l'IA d'un pays à l'autre. Plusieurs points se dégagent : les initiatives nationales en matière d'IA doivent tenir compte de la capacité de calcul ; les normes de collecte de données et de mesure nationales et régionales doivent être étendues ; les décideurs devraient connaître les besoins en capacité de calcul des systèmes d'IA ; les mesures portant spécifiquement sur l'IA devraient être différenciées des capacités de calcul à visée générale ; les travailleurs doivent avoir accès aux compétences et aux formations liées à la capacité de calcul pour l'IA ; et il convient d'analyser les chaînes d'approvisionnement et les intrants liés à la capacité de calcul pour l'IA.

Acknowledgements

This work is based on the findings of the OECD.AI Expert Group on AI Compute and Climate (hereafter the “Expert Group”) of the OECD Network of Experts, and the OECD Secretariat. It was prepared under the aegis of the OECD Working Party on Artificial Intelligence Governance (AIGO) and the OECD Committee for Digital Economy Policy (CDEP).

At the time of publishing, the Expert Group was co-chaired by Keith Strier (NVIDIA), Jack Clark (Anthropic) and Tamsin Heath (United Kingdom Department of Digital, Culture, Media and Sport). Jennifer Tyldesley (United Kingdom Department of Digital, Culture, Media and Sport), Sana Khareghani (former United Kingdom Office for AI) and Satoshi Matsuoka (RIKEN Centre for Computational Science, Japan) were formerly co-chairs. Jonathan Frankle (MosaicML), Arti Garg (Hewlett Packard Enterprise) and David Kanter (MLCommons) provided support and guidance to the Expert Group and co-chairs. Celine Caira, OECD Digital Economy Policy Division led the report development and drafting with contribution from Lennart Heim, AI Governance researcher at the Centre for the Governance of AI (GovAI), and the input of Karine Perset, OECD Digital Economy Policy Division. Gallia Daor, Dirk Pilat, Audrey Plonk, and Andrew Wyckoff, OECD, provided advice and oversight.

The paper benefited significantly from the contributions of those associated with the Expert Group, including input and review by Juan Manuel Ahuactzin (ProMagnus Company), Luis Aranda (OECD), Leonidas Aristodemou (OECD), Aviv Balasiano (Technology Infrastructure in the Israeli Innovation Authority), Gregg Barrett (Cirrus AI), Arnaud Bertrand (ATOS), Pascal Bouvry (LuxProvide), Eliana Cardoso Emediato de Azabuja (Ministry of Science, Technology and Innovation), Landon Davidson (NVIDIA), David Elison (Lenovo), Maria Jose Escobar Silva (Universidad Técnica Federico Santa María), Rebeca Escobar (Federal Telecommunications Institute), Liliana Fernández Gómez (Digital Development Directorate - National Planning Department), Nicole Formica-Schiller (German AI Association, KI-Bundesverband), Garth Gibson (Vector Institute for AI), Alexia González Fanfalone (OECD), Cyrus Hodes (World Climate Tech Summit), Taras Holoyad (Federal Network Agency for Electricity, Gas, Telecommunications, Post and Railway), Chen Hui (Infocomm and Media Development Authority), Vijay Janapa Redi (Harvard University John A. Paulson School of Engineering and Applied Sciences), Jan Jona Javoršek (Jožef Stefan Institute), Suzette Kent (Kent Advisory Services), Johannes Leon Kirnberger (Consultant on AI and Climate), Roland Krüppel (Federal Ministry for Education and Research), Jiwon Lee (Ministry of Technological Innovation and Digital Transition), Drew Lohn (Georgetown University Center for Security and Emerging Technology), Sasha Luccioni (Hugging Face), Angus Macoustra (Commonwealth Scientific and Industrial Research Organisation), Utpal Mangla (IBM), Ulrike Moetzel (Federal Ministry for Digital and Transport), Lorenzo Moretti (Ministry of Technological Innovation and Digital Transition), María Paula Mujica (High Presidential Advisory Office), Alistair Nolan (OECD), Marc-Etienne Ouimette (Amazon Web Services), Manish Parashar (National Science Foundation), Lynne Parker (United States Administration), Sally Radwan (UN Environment Programme), Anand Rao (PwC), Ghilaine Roquet (Digital Research Alliance of Canada), José Gustavo Sampaio Gontijo (Department of Digital Science, Technology and Innovation), Jayne Stancavage (Intel), Dimitris Stogiannis (National Documentation Centre), Lila Tretikov (Microsoft), Georgios Tritsaris (Sectoral Scientific Council in Natural Sciences (NCRTI, Greece), Bonis Vasilis (National Documentation Centre), Ott Velsberg (Ministry of

Economic Affairs), Verena Weber (OECD), Zee Kin Yeong (Infocomm Media Development Authority of Singapore), and Martin Zagler (Ministry of Industry, Business and Financial Affairs). The authors gratefully acknowledge the contributions made by individuals and institutions that took the time to participate in expert interviews and the contributions made by those who completed the online survey.

Finally, the authors thank Misha Pinkashov and John Tarver for editing this report, Jacqueline Lessoff for data visualisation assistance, and Denisa Bencze, Andreia Furtado, Angela Gosmann, Sebastian Ordelheide, Shellie Phillips and Sierra Wyllie for research, editorial and communications support, the overall quality of the report benefited significantly from their engagement.

Acronyms and abbreviations

AI	Artificial intelligence
AWS	Amazon Web Services
CPU	Central processing unit
EV	Electric vehicle
FLOPS	Floating-point operations per second
GPAI	Global Partnership for Artificial Intelligence
GPU	Graphics processing unit
HPC	High-performance computing
ICT	Information and communication technology
IGO	Intergovernmental organisation
IoT	Internet-of-Things
IT	Information technology
kW	Kilowatt
ML	Machine learning
NLP	Natural language processing
NPU	Neural processing unit
OECD	Organisation for Economic Co-operation and Development
PFLOPS	Peta floating-point operations per second
R&D	Research and development
SaaS	Software as a service
SME	Small and medium-sized enterprise
TPU	Tensor processing unit
TFLOPS	Tera floating-point operations per second
VPA	Virtual personal assistant

Table of contents

FIGURES

INFOGRAPHICS

BOXES

1 Introduction

1.1. Objective of this work

Artificial intelligence (AI) is transforming economies and societies, bringing opportunities for increased economic productivity, inclusive growth, and breakthroughs in addressing global challenges. Understanding countries' capacity and readiness to embrace this fast-evolving transition is essential, including the availability of relevant infrastructure enabling computation for AI at scale.

The creation and use of AI relies on key elements, such as a skilled workforce, enabling public policies, regulations and legal frameworks, access to data, and sufficient computing resources – commonly referred to as "compute". For machine learning (ML) based AI systems, there are two key steps involved in their development and use that are enabled by compute: (1) training, meaning the creation or selection of models/algorithms and their calibration, and (2) inferencing, meaning using the AI system to determine an output. While other key enablers have received significant attention in policy circles, the hardware, software, and related compute infrastructure that make AI advances possible receive comparatively less attention.

Ensuring countries have sufficient AI compute to meet their needs is critical to capturing AI's full economic potential. Many countries developed AI plans without a full assessment of whether they have sufficient domestic AI compute to realise these goals. The development of standardised measures for AI compute remains a policy and data gap. Policy makers require accurate and reliable measures of AI compute and how much national capacity they have, to make better-informed decisions and reap the full benefits of AI. Greater understanding of AI compute and its relationship to the diffusion of AI across OECD and partner economies can improve implementation of national AI strategies and guide future policymaking and investment.

Governments committed to the first intergovernmental standard on AI in the 2019 OECD Principles on Artificial Intelligence, "fostering the development of, and access to, a digital ecosystem for trustworthy AI", including underlying infrastructure such as AI compute (OECD, 2019[3]). Absent a measurement framework to facilitate the analysis of national AI compute capacity, "AI-compute divides" could be left unchecked within countries, such as between the private sector and academia (Ahmed and Wahed, 2020[4]), and between countries, such as between developed and emerging economies. This could create gaps between those that have the resources to create the complex AI models that lead to competitive advantage, inclusive growth, and productivity gains in a global digital economy, and those that do not.

The OECD.AI Expert Group on AI Compute and Climate (the Expert Group) advances understanding and measurement of AI compute to help policy makers understand their AI compute needs and work towards addressing them (Box 1). The Expert Group assists the OECD in developing a framework for countries to assess their domestic AI compute capacity, to establish baselines and benchmarks to guide public policy and investment decisions. In doing so, it helps countries answer three fundamental questions: (1) How much AI compute does the country have? (2) How much AI compute does the country need (i.e., is it sufficient to support national AI strategy objectives)? (3) How does it compare to other countries?

This report is informed by the Expert Group and delivers the next steps identified in a scoping note presented in December 2021 to the OECD Committee on Digital Economy Policy (CDEP). The Expert Group undertook a stocktaking of existing indicators, proxies, frameworks, and metrics for measuring AI compute at the national or sectoral level (Annexes B and C). Complemented by a gap analysis, this stocktaking helps avoid duplication of efforts in developing a measurement framework for data collection.

Box 1. The OECD.AI Expert Group on AI Compute and Climate

The OECD Network of Experts on AI (ONE AI) provides policy, technical, and business input to OECD analysis and recommendations. As a multidisciplinary and multistakeholder group, it provides the OECD with an outward perspective on AI, also serving as a platform to share information with other international initiatives. ONE AI raises awareness about trustworthy AI and sustainability issues, and other policy initiatives, particularly where international co-operation is useful.

The OECD.AI Expert Group on AI Compute and Climate (the Expert Group) advances understanding of AI compute and helps countries build awareness and work towards closing "AI compute divides" within and between countries. The Expert Group provides actionable and user-friendly evidence on AI compute, including its environmental impacts. In doing so, it enables policy makers to evaluate current and future national AI compute needs and corresponding capacity.

An AI compute divide can manifest within countries between the private sector and academia, as private-sector actors often have greater resources and access to AI compute to advance their objectives. An AI compute divide can also manifest and worsen between countries, namely between advanced and emerging economies, if governments cannot make informed decisions about investments to fulfil their national AI plans. This opens a gap in countries' ability to compute the complex AI models that lead to productivity gains in a global digital economy.

The Expert Group supports policy makers and practitioners in developing tools and indicators measurable at national level and that enable sufficient geographic coverage for benchmarking. Recommendations resulting from its work strive to be comprehensive, accessible to technical and non-technical audiences, and dynamic and time-proof, allowing for evolution as compute hardware and software advance (e.g., faster processors, larger memory, next-generation networks, quantum computing, etc.).

The Expert Group is co-chaired by Keith Strier (Vice President of Worldwide AI Initiatives at NVIDIA), Jack Clark (Co-Founder of Anthropic), and Tamsin Heath (Deputy Director of Economic Security at the Department of Digital, Culture, Media and Sport, United Kingdom). Jennifer Tyldesley (Department of Digital, Culture, Media and Sport, United Kingdom), Sana Khareghani (former Head of the Office for AI, United Kingdom) and Satoshi Matsuoka (Director, RIKEN Centre for Computational Science, Japan) were formerly co-chairs. The Expert Group meets virtually every three to four weeks since April 2021.

Source: OECD.AI Expert Group on AI Compute and Climate. Members are listed in Annex E with biographies are available on OECD.AI

1.2. Methodology and limitations

The methodology guiding this analysis relies on mixed-methods research, using publicly available qualitative and quantitative data and academic literature, expert interviews, and a survey undertaken by the Expert Group in 2022. To identify appropriate measurement tools, the expert group has developed a working definition of AI compute in addition to outlining key questions and further considerations.

This research encountered limitations to conducting evidence-based analysis. First, standardised and validated data on AI compute is not widely available. As such, this analysis is based on existing, publicly available data and academic papers in addition to the expertise and input of the Expert Group. Second, the market for AI compute is concentrated among a handful of hardware, software, and cloud computing companies, which limits access to validated data and methodologies. National-level and customer-level data on the supply and demand of AI compute is difficult to access and, in some cases, viewed as commercially sensitive proprietary information. Collaboration with private and public sector actors to collect data will be essential for advancing measurement work.

Third, this report primarily considers compute needs for ML, which is driving much of the demand for AI compute. Other AI systems, such as symbolic AI systems, have been less compute-intensive since they do not include a training process. Fourth, the report does not consider compute needs for processing and cleaning data for AI model training, which occurs at earlier stages of AI training and use.

A survey targeting an audience with expertise or knowledge of AI compute was conducted to inform the report (Annex D). There were 118 complete responses. Further analysis could benefit from the active participation of government representatives, private sector entities, and academia in systematic data collection efforts. This could be considered in the next phase of the Expert Group's work.

2 Evolving trends in compute

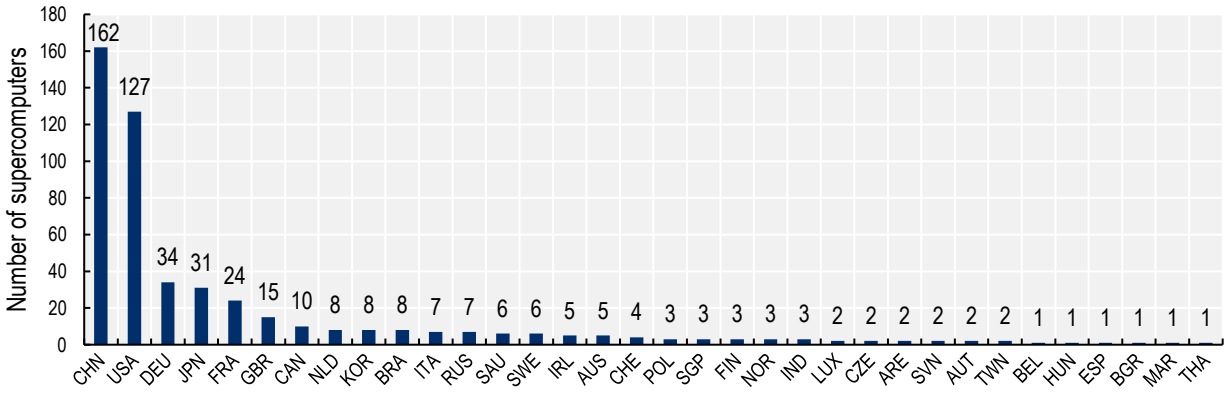
2.1. Trends in supercomputer performance

Few economies have supercomputers ranking as top computing systems, with emerging economies sparsely represented on the Top500 list

The Top500 list was created in 1993 to track the fastest supercomputers in the world primarily used for science. The Top500 methodology does not define "supercomputer", but instead uses a benchmark called Linpack to rank systems qualifying for the list. This means any supercomputer, regardless of its architecture, can make it into the Top500 list if it is able to solve a set of linear equations using floating point arithmetic. In recent years, supercomputer systems have been increasingly updated to also run AI-specific workloads, although the list does not distinguish supercomputers according to workload capacity specialised for AI. Analysis of the Top500 list can serve as a proxy measure to observe emerging or deepening compute divides between economies. As supercomputers increasingly are updated to also run AI-specific workloads, gaps could be observed between those having resources to create complex AI models leading to productivity gains, and those that do not.

The November 2022 Top500 list shows 34 economies with a "top supercomputer" according to the Top500 methodology (Figure 1). The highest concentration (32%) of top supercomputers is in the People's Republic of China (hereafter 'China'), followed by the United States (25%), Germany (7%), Japan (6%), France (5%) and the United Kingdom (3%) (Top500, 2022[5]). The 17 countries on the list from the European Union (EU27) make up a combined 21% of top supercomputers. Beyond this group, the rest of the world makes up 12% of top supercomputers. Nearly 90% of top supercomputers were developed in the last five years (Top500, 2022[5]). This highlights the speed with which hardware, infrastructure and software are being developed and brought to market.

Figure 1. Number of top supercomputers by economy according to the Top500, November 2022



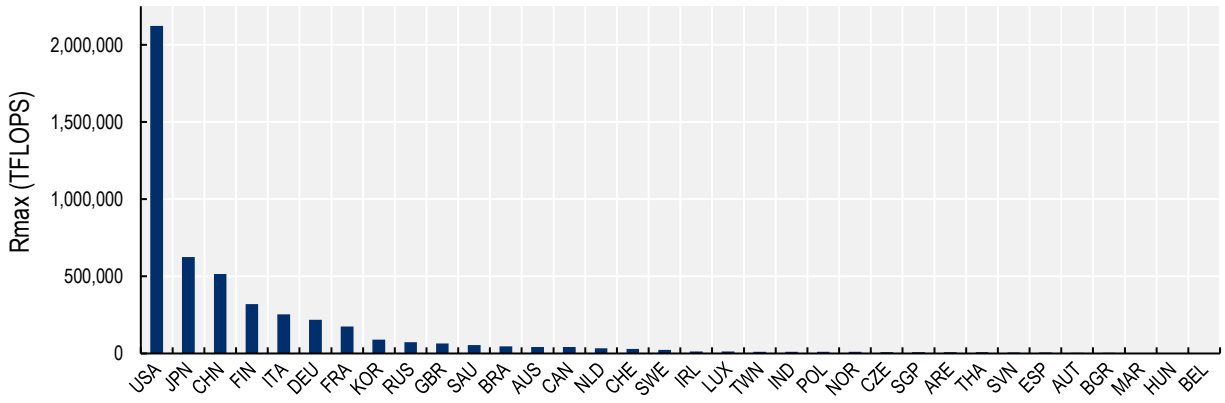
Note: The Top500 is released twice a year authored by Jack Dongarra, Martin Meuer, Horst Simon, and Erich Strohmaier. Contributions to the list are voluntary, posing methodological challenges. This Figure should be viewed as illustrative only and several caveats should be underlined. It does not consider the capacity of different supercomputers but the count of supercomputers by economy, (i.e., it treats different supercomputers as if they were the same, while significant variations in supercomputer capacity exist). It does not distinguish supercomputers according to workload capacity specialised for AI.
Source: Figure produced using data from the November 2022 Top500 list (Top500, 2022[5]).

Counting supercomputers does not give a full picture of national compute capacity as some supercomputers are more powerful than others

A simple count of Top500 list does not reveal the full picture of which economies hold the greatest supercomputer capacity, as this treats different supercomputers as if they were the same despite significant variations in supercomputer speed and performance. The Top500 ranks differences between supercomputer cores (processors), Rmax (a computer's maximum achieved performance), Rpeak (a computer's theoretical peak performance), and power (kW) using the Linpack benchmark. By analysing the November 2022 Top500 list, some researchers estimate that the performance of supercomputers has grown 630 times in terms of computational capacity since 2009[2] (Top500, 2022[5])

As of November 2022, the United States had five of the top 10 fastest ranked supercomputers on the list, including the first (called Frontier), while China had two of the top 10, followed by Japan, Finland, and Italy with one each. Analysis of the Top500 list by economy according to the sum of their maximum achieved performance (Rmax, measured in tera floating-point operations per second, or TFLOPS), shows that the United States has the highest share of total compute performance on the list (44%), followed by Japan (13%) and China (11%) (Figure 2). This shows that counting supercomputers does not give a full picture of national compute capacity, as some supercomputers are more powerful than others.

Figure 2. Top500 supercomputers by economy ranked by total Rmax, a computer's maximum achieved performance, November 2022



Note: This figure should be taken only as a preliminary and directional proxy metric for national compute capacity with the caveats outlined in Figure 1. In addition, as workloads cannot be run across multiple supercomputers, this measure should be viewed with limitations (e.g., 10 supercomputers that add up to the same sum of Rmax as a single supercomputer would not be equivalent).
Source: Figure produced using data from the November 2022 Top500 list (Top500, 2022[5])

2.2. Trends in compute for artificial intelligence

State-of-the-art AI systems increasingly depend on high-performance compute

Researchers estimate that the computational capabilities required to train modern ML systems, measured in floating-point operations per second (FLOPS), has grown by hundreds of thousands of times since 2012[1] (OpenAI, 2018[1]; Sevilla et al., 2022[2]), despite algorithmic and software improvements that reduce computing power needs. This is likely driven by the increasing capabilities of large, compute-intensive AI systems (Kaplan et al., 2020[6]; Hoffmann et al., 2022[7]). Research also notes that compute demands such as processing power for AI systems has grown faster than hardware performance, particularly for deep-learning applications like machine translation, object detection, and image classification (Thompson et al., 2020[8]).

AI compute is not well understood beyond specialised technical and policy communities

While awareness is growing of the importance of national policies for AI compute, its technical nature makes it less understood outside specialised technical and policy communities. Many private-sector actors have observed the growing reliance of AI systems on compute and made corresponding strategic investments. Companies providing cloud computing services leverage existing infrastructure to meet internal needs and serve customers, such as through infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS) cloud models. According to Eurostat, up to 41% of enterprises in the EU used some type of cloud computing in 2021 (Eurostat, 2021[9]). Examples include Google Cloud, Microsoft Azure, and Amazon Web Services (AWS), which provide cloud services enabling access to software applications, servers, storage, compute, and more, including for AI training and inference.

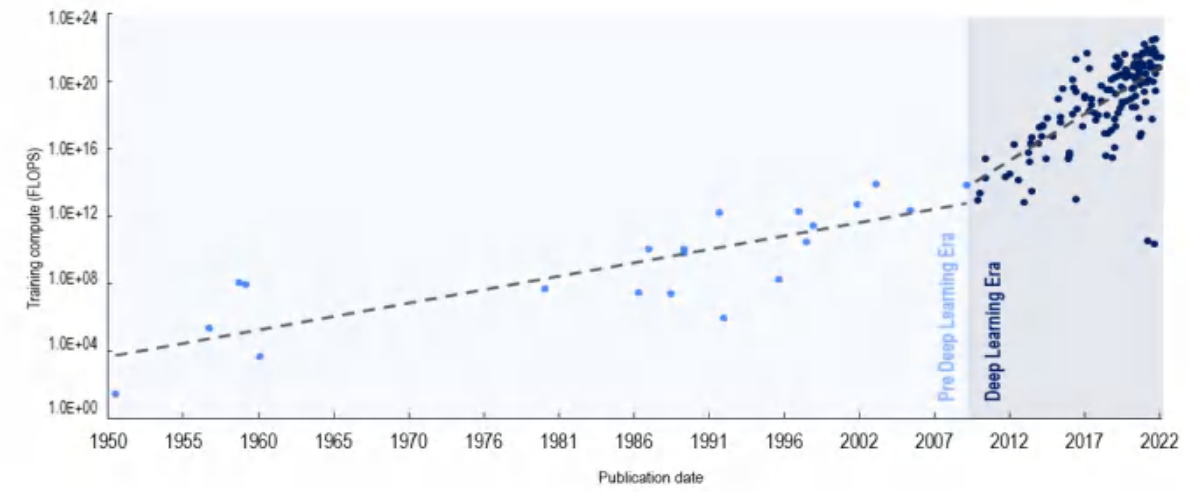
Securing specialised hardware for AI involves complex supply chains

Securing specialised infrastructure and hardware purpose-built for AI can be challenging due to complex supply chains, as illustrated by bottlenecks in the semiconductor industry (Khan, 2021[10]). Integrated circuits or computer chips made of semiconductors are the "brains of modern electronic equipment, storing information and performing the logic operations that enable devices such as smartphones, computers, and servers to operate" (OECD, 2019[11]). Any electronic device can have multiple integrated circuits fulfilling specific functions, such as CPUs or chips specifically designed for power management, memory, graphics, and more. Demands on semiconductor supply chains have grown in recent years, especially as digital and AI-enabled technologies become more commonplace, such as Internet-of-Things (IoT) devices, smart energy grids, and electric vehicles (EVs). The semiconductor supply chain is also highly concentrated, making it more vulnerable to shocks (OECD, 2019[11]).

The prominence of deep learning dramatically increased the size of machine learning systems and their compute demands

Starting in about 2010, the prominence of deep learning dramatically increased the size of ML systems and their compute demands (Figure 3). Satisfying this demand was partially enabled by transitioning from general-purpose processors, such as Central Processing Units (CPUs), to processors that include specialised hardware and support more efficient compute execution for certain operations (i.e., requiring less energy and more computations per unit time). Today, ML systems are predominantly trained on specialised processors that comprise hardware optimised for certain types of operations, such as Graphics Processing Units (GPUs), Tensor Processing Units (TPUs), Neural Processing Units (NPUs), and others. Training ML systems on general-purpose hardware is less efficient. In recent years, interest has grown significantly among governments and private sector actors in increasing and securing supply chains for such specialised hardware (Khan, 2020[9]).

Figure 3. Estimated compute used for training milestone ML systems between 1952-2022

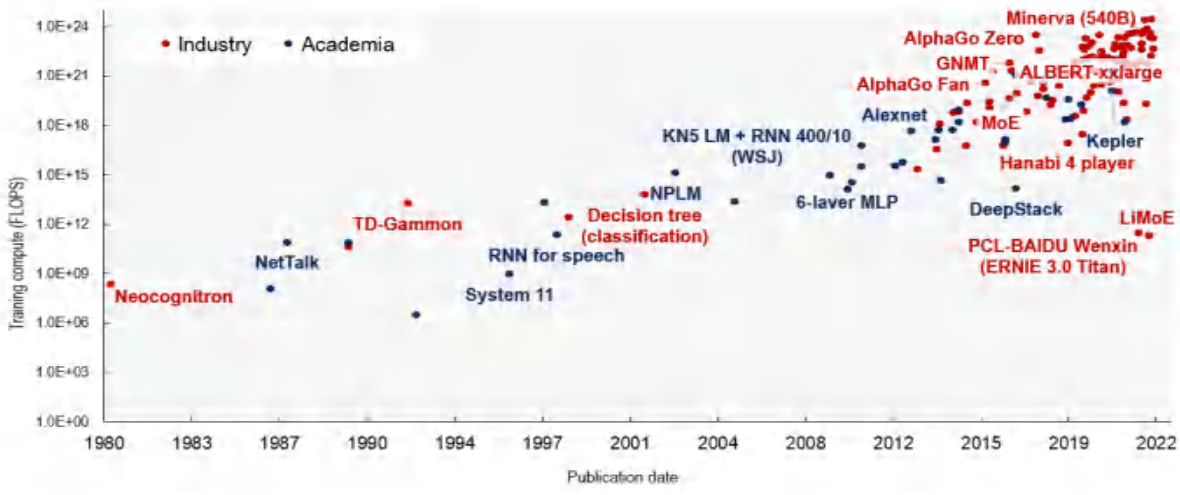


Source: Figure produced and adapted from data included in original work by (Sevilla et al., 2022[2])

Industry is training an increasing number of large AI models compared to academia

A compute divide can emerge and worsen between the public and private sectors because, increasingly, public sector entities do not have the resources to train cutting edge AI models. Industry, rather than academia, is increasingly providing and using the compute capacity and specialised labour required for state-of-the-art ML research and training large AI models (Figure 4) (Ahmed and Wahed, 2020[4]; Ganguli et al., 2022[13]; Sevilla et al., 2022[2]). Several countries announced initiatives to increase the compute available for research and academia, including the United States National AI Research Resource (NAIRR) and Canada's Digital Research Infrastructure Strategy, in addition to initiatives to take stock of compute capacity and needs, including for researchers, such as the Canadian Digital Research Infrastructure Needs Assessment and the United Kingdom's 2022 Future of Compute review. Section 5 discusses additional national AI initiatives related to compute.

Figure 4. Estimated compute used for training milestone ML systems classified by compute provider (industry or academia) between 1980-2022



Note: According to Sevilla et al., 2022, "Sector is based on affiliation of the research paper authors." and "If the authors had affiliations in both Academia and Industry, the sector was labelled Industry because Industry-controlled computation is preferred in practice."
Source: Figure produced and adapted from data included in original work by (Sevilla et al., 2022[2])

3 Measuring AI compute: Definitions, scoping considerations, and measurement challenges

3.1. AI compute: What is it and what is it for?

This section outlines discussions on what artificial intelligence (AI) compute is and does. According to the 2019 OECD Recommendation of the Council on AI [OECD/LEGAL/0449], AI is defined as "a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments". While AI can be perceived as an intangible, technical system, it is grounded in physical infrastructure and hardware, which is increasingly specialised for AI development and use.

In its scoping work, the OECD.AI Expert Group on AI Compute and Climate (the Expert Group) proposed a working definition of AI compute understandable to technical and policy communities (Box 2). The Expert Group found that, while there is no standard definition of AI compute (Annex A), its core elements are understood by technical AI experts, developers, and practitioners. The Expert Group thus proposes defining AI compute as "one or more stacks of hardware and software used to support specialised AI workloads and applications in an efficient manner" with requirements varying significantly according to the user's needs. This definition results from discussions by experts from the group and beyond and might be further refined.

This report uses several terms related to AI compute. "Computing resources" or simply "compute" refer to general-purpose compute, which is not necessarily purpose built for AI applications such as AI training or inference. "AI computing resources" or "AI compute" refer to the physical hardware and software infrastructure supporting AI workloads, including one or more "stacks" (layers) of hardware and software used to support specialised AI workloads and applications in an efficient manner. "National AI compute capacity" means the totality of resources that can be used to support AI development and use towards achieving national policy goals.

AI compute covers a range of different technologies, from chips to data servers to cloud computing. AI compute enables AI systems' training, meaning the creation or selection of models/algorithms and their calibration, and inferencing, or using the AI system to determine an output. This results in AI compute requirements varying significantly according to user needs. AI compute can be located at and accessed in several ways:

- **Centrally in data centres,** as infrastructure in physical facilities that house the computational hardware, networking equipment, software, and data used for AI.
- **Centrally in the cloud,** as a service through public or private cloud networks.
- **At the edge on decentralised devices,** contained directly on stand-alone, end-use devices for local AI inferencing, for instance on mobile, Internet-of-Things (IoT) devices.

Box 2. Defining and scoping AI compute

Between April 2021 and April 2022, the Expert Group conducted eight meetings, interviews with more than 25 experts, and a survey to inform work on defining and scoping AI compute:

> *"AI computing resources ('AI compute') include one or more stacks of hardware and software used to support specialised AI workloads and applications in an efficient manner."*

This definition highlights several properties central to a common understanding of AI compute:

- **AI compute includes stacks of hardware and software.** AI workloads are not performed by one hardware or software component, but by one or more "stacks" (layers) of components. The stacks include storage, memory, networking infrastructure, and more, designed to support AI-specific workloads and applications that run mathematical calculations and process data at scale. Efficient interaction between the hardware and software stacks is crucial for AI compute.
- **AI compute stacks are specialised for AI workloads.** Specialised hardware enables AI training and use. For example, graphics processing units (GPUs) are purpose-built for highly parallelised computing, in which many calculations are carried out simultaneously, making them highly efficient for certain AI model types, such as deep learning. AI compute stacks are becoming increasingly specialised, as AI applications, the number of parameters, and dataset sizes continue to grow.
- **AI compute requirements can vary significantly.** Depending on the application, AI system lifecycle stage, and size of the system, the AI compute needed can vary from large, high-performance computing clusters or compute hyperscale cloud providers to smaller data-science laptops and workstations. Consequently, compute requirements vary significantly based on national AI plans and along the AI system lifecycle.
- **AI compute supports AI workloads and applications in an efficient manner.** AI compute differs from general-purpose compute in that it can support AI workloads and applications in an efficient manner, such as through optimised execution time and energy usage. This efficiency is critical for conducting AI R&D, using large models and datasets.

Source: OECD.AI Expert Group on AI Compute and Climate

To understand the role of compute in AI systems, it is also important to understand the basic AI production function, described by three enablers: algorithms, data, and compute (Figure 5). Compute is a substantial component of AI systems and a driver of their improved capabilities over time. It is distinguished from data and algorithms by being grounded in "stacks" (layers) of physical infrastructure and hardware, along with software specialised for AI. Such stacks, made up of a variety of hardware and software components and configurations, are part of why AI compute is difficult to quantify. While the compute needs of AI systems and the specifications of hardware can be estimated, defining an "all-encompassing unit of AI compute" has not been possible due its complexity.

Figure 5. Examples of AI compute enablers

Source: OECD.AI Expert Group on AI Compute and Climate

In addition, compute often requires significant natural resources, including energy and mineral demands for hardware production, and energy and water consumption during operation. This is explored in a parallel report informed by the Expert Group, in collaboration with experts from the Responsible AI Working Group of the Global Partnership on AI (GPAI) (OECD, 2022[14]).

Compute requirements can vary significantly for an AI system depending on its lifecycle stage. The OECD defines an AI system lifecycle as encompassing the following phases: (1) plan and design; (2) collect and process data; (3) build and use the model; (4) verify and validate the model; (5) deploy; and (6) operate and monitor the system (OECD, 2022[15]). For machine learning (ML) systems, two lifecycle phases stand out for their compute needs: training (building the AI system) and inferencing (operation).

Training an AI model such as a neural network – one of the most computationally intensive types of AI models – involves determining the value of weights and biases (also referred to simply as "learning") from data presented to the system. This is a fundamental component of ML, regardless of whether supervised, unsupervised, or reinforcement learning is used. Once a neural network is trained, it generates the output through a computational process applying the trained weights against new input data. This is referred to as inferencing (or "forward pass"). Once trained, a network can be distributed and deployed for application. At this point, the network is mostly static: all computations and intermediate steps are defined, and only an input (such as an image to classify) is necessary to carry out inferencing. Examples of inferencing include looking up information using a search engine (e.g., a single Google search) or talking to a virtual personal assistant (VPA) (e.g., Siri, Alexa, or others).

A complete *training* run is computationally more intensive than compute used to make a single *inferenc*e (Bengio, Courville and Goodfellow, 2016[16]). There are two primary reasons for this. First, training of the weights is iterative: many cycles are required for a single input to obtain the desired result. Second (at least for supervised learning), training data needs to be available to the compute system, which requires memory capacity. Due to the combination of these two factors, training is thus usually a more complex process in terms of memory and compute resources. Given the significant data and compute requirements, training is more likely to be conducted on centralised, high-performance computers. In contrast, AI deployment (i.e., running inferences) is more variable regarding AI compute requirements. Inferencing can be conducted on computationally less powerful devices, such as smartphones, for instance at the edge (e.g., using IoT).

However, while a single *training* run is more computationally intensive than a single *inference*, the inferencing stage overall typically requires more compute in an AI system's lifecycle because ML systems are usually trained only a few times during their development phase, whereas inferencing is executed repeatedly every time a system is used during the lifetime of its deployment (Patterson et al., 2021[17]; OECD, 2022[15]; Bengio, Courville and Goodfellow, 2016[16]).

The compute used for training versus inference can also be impacted by whether the model is using techniques such as transfer learning or federated learning. *Transfer learning* allows for some efficiency gains through the "re-training" of models. For example, a model trained for image recognition generally can be re-trained to recognise specific images, for instance images of cats. This can enable efficiency as pre-trained models can be repurposed for specific purposes. Another example is *federated learning*, a ML technique that conducts training across multiple decentralised servers or edge devices holding specific data (e.g., using IoT), without exchanging them, which can be used to train models more efficiently in some cases. Both are examples of how measuring the relationship between compute needed for training and inference can be dynamic and depend on the model and task at hand.

3.2. Measurement challenges

Measuring AI compute capacity and needs is particularly challenging. At present, very few tools and indicators exist to measure AI compute. Literature on AI compute typically focuses on the performance

measurement of compute systems, such as application performance benchmarks like MLPerf or throughput benchmarks like the Top500 list. Other methods use the number of mathematical calculations a computer can complete in a second (floating-point operations per second, or FLOPS) as an indicator of compute performance. While measures of compute performance are useful, they are not a complete indicator of collective national compute capacity nor of a country's AI compute needs.

What qualifies as "domestic" AI compute may vary by country, for example being subject to domestic laws and regulations and physically located within a national jurisdiction. Policy makers will need to consider whether AI compute can be classified as domestic if it is (1) owned and operated by a non-domestic private or public sector actor and/or (2) physically located in another country. Aggregating the performance of individual AI systems within a country could be one way to calculate national AI compute capacity, but this approach has limitations. Commonly used benchmarks are narrowly formulated to define performance under very strict conditions (e.g., the Linpack benchmark) and might not be applicable to all AI systems in a country. Another approach is to count the number of discrete AI systems and group them by "class" of performance, such as leadership-class AI systems and centre-class AI systems. This might provide less-specific results but is more user-friendly.

Another measurement challenge is that compute can be general-purpose, meaning that compute infrastructure can be used for AI workloads and non-AI workloads, such as mathematical and scientific modelling and other compute needs not directly related to AI. This challenge is particularly relevant to hardware and infrastructure as data centres and high-performance computing (HPC) infrastructure can have a variety of applications in addition to AI. Few estimates of AI-specific workloads exist, with these rarely differentiating between AI training and use. According to one study by Google, its overall energy use for ML workloads consistently represented less than 15% of total energy use over 2019-21 (Patterson, 2022[58]). Other estimates use customer spending to approximate the percentage of compute used between AI training and inference workloads. For example, a large cloud compute provider[3] estimates that its enterprise customers spend 7-10% of their total compute infrastructure expenditure on supporting AI and ML applications, broken down to 3-4.5% for training and 4-4.5% for inference. This includes about 60% spent on compute platforms featuring hardware accelerators like GPUs and about 40% spent on CPU-based compute platforms. Such numbers can inform estimates of AI-specific use while shedding light on how impacts differ according to whether compute is used for AI training or inference.

The Expert Group focuses on creating a measurement framework for AI compute at the national level, which also poses specific challenges. Countries participate in a variety of international and regional initiatives like research collaborations on HPC, which complicates assigning AI compute capacity to individual countries. National capacity accessed through the cloud raises the same issue as compute accessed domestically through the cloud could rely on servers and data centres located across borders and in different jurisdictions.

Determining skills and job titles related to AI compute activities is also a challenge. The 2008 International Standard Classification of Occupations (ISCO-08) and many national occupation classifications do not distinguish AI compute specialised occupations from general software and ICT development, manufacturing, and maintenance jobs (United States Census Bureau, 2022[19]; International Labour Organization, 2016[20]). This makes international comparability challenging, especially when AI compute related job titles are poorly defined. For example, a "data scientist" job posting might ask for skills in AI modelling, training optimisation for hardware, big data, and various AI domains (e.g., NLP or computer vision). These skills overlap with job titles like "machine learning specialist" and "data engineer", with some being even more specific, like "computer vision specialist". The skills listed in AI-related job postings also differ by country due to differing national technology environments and demands for experience, such as managerial skills (Samek, Squicciarini and Cammeraat, 2021[21]).

3.3. Insights from preliminary survey results

The preliminary results of the public survey on AI compute launched by the Expert Group highlight some of these measurement challenges (Annex D). Of respondents, 27% reported that they measure AI compute capacity, 22% reported that they use some metrics but not regularly, and 31% reported that they do not measure how much AI compute they have and do not have metrics and measurement tools in place (Figure D.5). In contrast, 20% reported that they did not have sufficient information to answer this question and that they did not know whether they measure AI compute. Furthermore, 52% of respondents reported challenges accessing sufficient AI compute, compared to 30% reporting no challenges and 18% reporting that they did not know whether they had challenges (Figure D.6).

When asked about the top barriers or challenges to accessing AI compute, 44% of respondents cited the cost of AI compute, followed by expertise (20%), availability (13%), and suitability (5%) (Figure D.7). This highlights cost as an important factor for planning effective use of AI compute and access. When asked about the percentage of their organisation's total annual costs spent on AI compute, 37% reported that they did not know, 5% reported no annual costs spent on AI compute, 26% reported 10-40% of costs, and 3% reported that AI compute costs were 50% or more of annual costs (Figure D.8).

4 Blueprint for developing a national AI compute plan

4.1. Aligning compute capacity with national AI strategies

Many countries have produced national AI strategies without explicit consideration of whether they have the corresponding infrastructure, hardware, and skilled labour to execute such plans and achieve national artificial intelligence (AI) policy goals. To address this gap, the Expert Group developed considerations to help policy makers align national compute capacity and future investments with national AI strategies. These considerations are not exhaustive and vary according to national contexts and AI needs. They are the outcome of extensive discussions with Expert Group members and offer an overview of the current thinking on how to measure and plan national AI compute capacity for current and future needs.

Policy makers should consider AI compute investments relative to national policy objectives, including public-sector budget allocations and private-sector investments. Policy makers should recognise that there are different ways to boost domestic AI compute capacity and the most resilient approach will depend on a country's context and needs. Such an approach could include investments in nationally owned or sponsored AI supercomputers and/or strategic partnerships with global and regional commercial cloud providers. But valuable AI compute can also be small, especially for students and junior researchers. Policy makers can keep in mind that even a data science laptop or workstation, which do not require the overhead costs of a data centre, can be a powerful vehicle to AI innovation, broadening access and helping to close compute divides.

Policy makers also should consider how public- and private-sector investments in domestic AI compute capacity can advance different types of policy objectives. For example, *scaling up* AI compute involves investment in a smaller number of larger AI systems for training the largest and most complex AI models (e.g., supporting advances in domains such as natural-language processing (NLP), precision medicine, and autonomous vehicle development). Alternately, *scaling out* AI compute involves investment in a larger number of smaller AI systems to enable AI R&D projects such as workforce training and student education (e.g., where the goal is more about access than breakthroughs). The scaling out approach is commonly seen in countries such as Thailand and Indonesia, where multiple smaller AI clusters are installed in universities with government support to broaden access. These examples of AI-related policy goals and their implications for AI compute are summarised below:

- **Scale-up AI policy** develops and uses AI to achieve cutting-edge innovation in specific domains (e.g., health, transport, agriculture) to solve complex problems and increase or maintain a country's competitiveness in that domain.
- **Scale-out AI policy** promotes AI diffusion across sectors of the economy to unlock productivity gains and innovation at scale. It typically promotes inclusion and aims to produce AI benefits that are widely shared.

Policy makers might wish to conduct a needs assessment by developing an AI compute country profile. Some initial contextual factors and country profiles are presented in Box 3.

Box 3. Sample considerations for national AI compute profiles

Preliminary contextual factors to consider:

- Economic development level
- Telecommunications network maturity
- National AI strategies and private sector AI needs (e.g., quality and availability of infrastructure)
- Level of digital adoption (e.g., in private and public sectors)
- Availability of AI inputs (e.g., data maturity, prevalence of AI-ready datasets)
- Workforce AI literacy (e.g., in private and public sectors)
- Geography and access to supply chains (e.g., space and capacity to build data centres and HPC clusters within them)

Country profile #1: Emerging economy

A policy maker in an emerging economy that has a mature telecommunications system, limited in-country data centre capacity, and a low level of science, technology, engineering, and math (STEM) education might wish to:

- Understand how to best leverage the global supply of AI compute within the country
- Explore how partnerships and capacity-building programs can build infrastructure, help train the workforce, and grow the level of AI literacy (e.g., using STEM education)
- Plan a compute strategy to build a baseline of compute capacity to stimulate economic growth

Country profile #2: Advanced economy

A policy maker in an advanced economy with a mature telecommunications system, a large amount of in-country data centre capacity (including some local hyperscale data centres), and a high level of STEM education might wish to:

- Understand options to invest in AI compute to utilise the existing high level of STEM education
- Analyse how accessible existing AI compute is for local businesses
- Plan a compute strategy to double down on existing investments and strive for long-term gains in economic competitiveness

Note: This offers illustrative thoughts on elements that can impact national AI compute profiles. This framework could be further developed and adapted to fit specific national needs and contexts.
Source: OECD.AI Expert Group on AI Compute and Climate

4.2. Considerations for a national AI compute plan

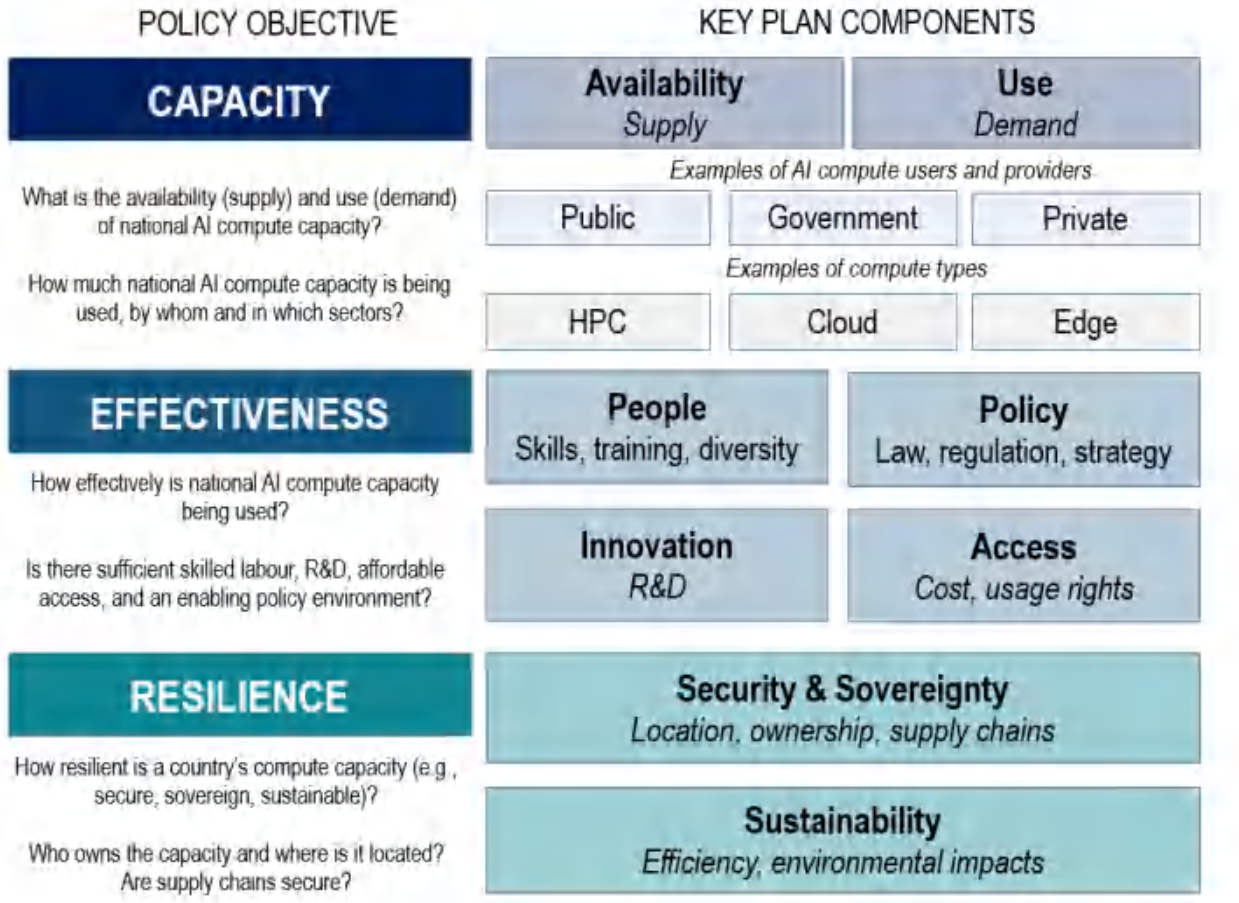
This section offers a blueprint for the creation of a national AI compute plan and describes considerations for national policy makers and practitioners (Figure 6). Alongside implementation of a national AI compute plan, policy makers should develop indicators and metrics to evaluate its success and inform future evidence-based policy. It can also be used as a basis for policy makers to develop measurement and evaluation frameworks and begin collecting data on AI compute. The Expert Group proposes this blueprint with accompanying questions for policy makers to tailor and guide AI compute capacity investments to meet national AI ecosystem needs. Each plan component and consideration is described in more detail below.

A national AI compute plan should align with existing national AI strategies and centre around three fundamental questions:

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

To answer these questions, policy makers can consider three overarching categories as part of a national AI compute plan – capacity, effectiveness, and resilience – which include subcomponents and can be used to develop metrics and indicators for evaluation (Figure 6). Each of these components are presented in more detail below along with questions for policy makers to consider.

Figure 6. Blueprint for national AI compute plans



Source: OECD.AI Expert Group on AI Compute and Climate

Capacity

Measuring a country's national capacity for AI compute is challenging. It involves developing baseline supply and demand measures and forecasting to ensure investments reflect future needs and the fast-changing pace of AI (Figure 7).

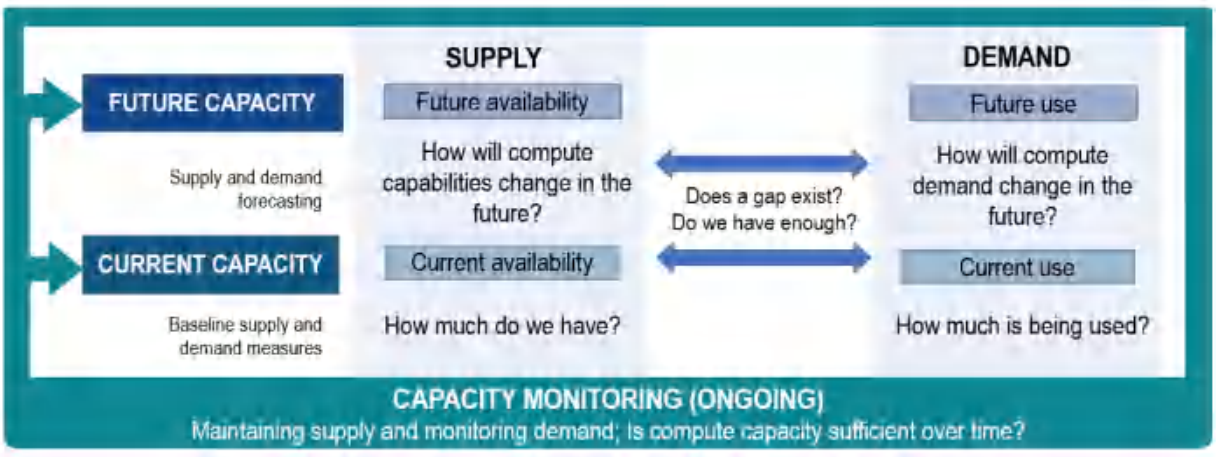
Figure 7. Policy objectives and considerations for AI compute capacity

Source: OECD.AI Expert Group on AI Compute and Climate

To help policy makers address these measurement challenges, the Expert Group developed a framework for measuring current and future national AI compute capacity and ensuring ongoing capacity monitoring (Figure 8). A preliminary list of measurement indicators under discussion is proposed in Annex C, which map to the components of the blueprint.

1) **Measuring *current* AI compute capacity and needs** by taking stock of national supply and demand for AI compute
2) **Estimating *future* AI compute capacity and needs** by anticipating advances in technical compute capabilities against future demands and ambitions in national AI plans
3) **Ongoing monitoring of AI compute supply and demand** by maintaining supply and tracking demand for national AI compute over time

Figure 8. Framework for measuring national AI compute capacity and ensuring ongoing monitoring



Note: This is a preliminary framework which could continue to evolve as the Expert Group continues its work.
Source: OECD.AI Expert Group on AI Compute and Climate

Measuring current AI compute capacity and needs

Current supply: How much do we have?

A first step in measuring a country's AI compute capacity is taking stock of the AI compute supply currently available in that country. This could help answer the question, "How much do we have?" The following questions can be relevant to measuring compute supply:

- How much general-purpose compute exists nationally and how much can be used for AI? How relevant is the compute capacity for AI?
- How much AI compute supply exists across central compute (i.e., on-premise HPC clusters), cloud compute, and edge compute (i.e., connected devices) resources?
- What is the landscape of AI compute providers in a country? What is the breakdown by private, public and government-owned providers? Are there partnerships with non-domestic cloud providers?
- What should be counted as national AI compute capacity considering the various locations of AI compute, such as cloud compute servers in another country?
- Is there a difference between how much AI compute supply exists and how much is available? What are the key factors influencing the availability of the supply of AI compute?
- How do cost considerations impact the supply of AI compute?

Current demand: How much do we need?

In addition to estimating the national supply of AI compute, estimates of the compute demand by various actors and types of AI systems are important to inform projects and decisions, especially considering the rapid technological advances of AI in recent years. The following questions can be relevant to assessing compute demand of AI systems:

- Which actors (e.g., private, public, government sectors) are using and need AI compute?
- How do AI compute needs differ across industry and research sectors? What specific needs for specialised compute do specific sectors have?
- How do AI compute needs differ across types of AI, such as symbolic AI or ML?
- How do AI compute needs differ across AI applications such as NLP, speech recognition, computer vision, generation of recommendations, generative AI and other applications?
- How do compute needs differ along the AI system lifecycle (e.g., from training to inferencing)?
- How does demand for specialised AI compute skills vary by sector?
- How do cost considerations impact the demand for AI compute?

Measuring future AI compute capacity and needs

Future supply: How will compute capabilities change in the future?

Growth trends of AI compute hardware, software, and other resources impact supply projections. The following questions can help estimate future supply of AI compute based on technical advances:

- What do technologists, foresight specialists, and others forecast as the next advances in computing hardware and infrastructure for AI?
- How are AI compute technical capabilities expected to evolve with the introduction of more efficient hardware (e.g., quantum computing and networks like 5G, 6G, etc.), software and other compute resources?

- How are AI compute technical advances expected to compliment and interact with existing national infrastructure for AI compute?
- Which skills might be needed to effectively leverage advances in AI compute?

Future demand: How will compute demand change in the future?

Few national AI strategies and associated action plans anticipate development in demand for AI compute. Forecasting AI compute demand can be challenging as AI applications grow as a general-purpose technology. The following questions can help policy makers assess future demand for AI compute as part of national AI ambitions:

- Which emerging sectors or AI applications requiring significant or specialised compute could be of national economic or scientific importance?
- How might compute demand for different types of AI (e.g., symbolic AI or ML) evolve?

Capacity monitoring (ongoing)

Monitoring demand and maintaining supply for AI compute is key as AI compute sufficiency requires constant re-evaluation considering new applications and investments. The following questions could be considered:

- Which strategic investments could secure adequate ongoing supply of compute for national AI needs?
- Is current national AI compute capacity dynamic enough to adapt to new technological advances and applications? What share of current systems can be easily and affordably upgraded or extended?
- What share of current AI compute resources could become outdated or obsolete?
- How resilient is current national AI compute capacity to supply-side shocks (i.e., fragility of supply chains), natural disasters, and geopolitical considerations?

Effectiveness

In addition to taking stock of the availability and use of AI compute in a country, it is important to consider whether a compute divide exists due to ineffective use of compute. For example, a lack of skilled labour, innovation and R&D ecosystems, enabling laws and regulations, as well as high costs and other barriers to accessing AI compute can cause even state-of-the-art infrastructure to be used ineffectively.

Figure 9. Policy objectives and considerations for AI compute effectiveness

Source: OECD.AI Expert Group on AI Compute and Climate

People (skills, training, diversity)

Specialised skills, often engineers or those with technical hardware expertise, are needed to use AI compute resources effectively. Preliminary results from the survey on AI compute (Annex D) highlight this point. Most survey respondents reported at least one full-time equivalent (FTE) worker dedicated to the management and use of AI computing resources (Figure D.4), and over 10% of respondents reported 250 or more FTE workers dedicated to this work. Only 12% of respondents reported zero FTE workers, and around 16% reported that they did not know.

With the increasing demand of AI workloads, skilled labour might be a bottleneck to deployment of AI compute. Expertise to ensure that the configuration of hardware and software stack(s) is efficiently deployed and easy to use is critical to enabling effective compute capacity. Perspectives from diverse disciplines and backgrounds are also critical to close compute divides between developed and emerging economies, and between public, academic, and private sector organisations. The Expert Group is examining these challenges and how skills for AI compute differ from AI skills more broadly. The following questions could be considered:

- Is there sufficient supply of talent nationally with the skills to enable the effective use of AI compute?
- What skills are required for the effective use of AI compute? How do these differ from AI skills in general?
- What is the demand for and prevalence of these skills nationally? Are there skills shortages? Can countries attract these skills?
- Are perspectives from diverse disciplines and backgrounds being considered in the planning and implementation of national AI compute plans?
- Are domestic AI education and training programs promoting trustworthy AI principles in learning about effectively using AI compute?

Policy (law, regulation, strategy)

National policy environments that facilitate the effective use of compute infrastructure play a foundational role in successful AI compute plans. The laws, regulations, and strategies surrounding governance of and access to compute are critical to its effective use. Countries and regions take different approaches to governing the digital infrastructure required for AI development and use, from national HPC or cloud resource initiatives, to targeted hiring and skills policies. The following questions could be considered:

- What laws and regulations govern national compute capacity and are they fit to serve today's innovation economy, national context, and the needs of AI systems?
- Are there laws and regulations that create red tape and other undue administrative burdens on users and providers of AI compute? How does this vary between the public and private sectors?
- Have policies that involve partnership with non-domestic cloud providers been considered?
- Have policies been considered to subsidise high-powered data-science laptops for AI developers, researchers, and students to close divides between private- and public-sector compute availability?
- What can be learned from countries that leverage AI compute capacity to produce breakthroughs and increase domestic competitiveness?

Innovation (Research and Development)

Research and development (R&D) support innovation and advances in AI compute infrastructure and stack architectures, enabling significant efficiency gains and breakthroughs in AI discoveries. Such research and innovation drive technology advances that can influence the investment decisions countries make

regarding compute infrastructure and hardware to compete in a global digital economy. The following questions could be considered:

- What compute technology advancements impacted the domestic AI ecosystem in the last decade?
- Which types of companies (i.e., private or public sector) fund and conduct R&D for breakthroughs in AI compute technology?
- How do new technologies become available and/or are they open source?
- How are innovation and R&D advances changing the skills needed to adopt new technologies? Do national training programs require updating?

Access (cost, usage rights)

Ways of accessing compute include renting cloud compute from private companies, accessing compute directly on-premises through data centres, or accessing compute through research collaborations and public-private partnerships. Barriers to access include lack of awareness, service reliability, and expertise, as well as high cost. Who owns the compute capacity can also impact the ways and ease with which capacity is accessed according to usage rights. The following questions could be considered:

- How much AI compute is accessible across public, private, and academic ownership models? Can AI be owned, operated, and made accessible by governments, universities, or the private sector, for instance by renting cloud compute?
- How do compute needs differ according to varying means of access and usage rights? Is there increasing demand for access to cloud computing resources?
- How can compute capacity be measured in the cloud, given access models that cross jurisdictions?
- How do usage rights impede access to compute capacity for different groups (i.e., public sector, universities, research institutes, private sector companies)?
- How do cost considerations impact the supply of AI compute? Is cost a barrier to investment?

Resilience

Resilience considerations include concerns related to security and sovereignty, such as location, ownership and supply chains, and to environmental sustainability.

Figure 10. Policy objectives and considerations for AI compute resilience

Source: OECD.AI Expert Group on AI Compute and Climate

Security and sovereignty (location, ownership, supply chains)

General sovereignty and security

A country's AI compute capacity can be located domestically or internationally, such as in a data centre located within the country's borders or one abroad. Such AI compute capacity could be considered "sovereign" if it is subject to domestic laws and regulations. Whether investments improve AI compute capacity domestically or abroad depends on national goals and could be linked to national security and privacy objectives, among others. Countries could also have infrastructure reliability and cybersecurity concerns, including whether national electricity grids have sufficient capacity to support desired national AI compute plans and whether AI compute infrastructure is secure from malicious activities like cyberattacks. The following questions could be considered:

- How much and what type of AI compute capacity sits within a country's borders and is governed by national laws versus in other jurisdictions?
- What are the trade-offs in sourcing AI compute domestically or abroad?
- How can domestic AI compute be distinguished from internationally available compute, for instance if accessed through commercial cloud providers?
- How do national security considerations and privacy objectives intersect with AI compute plans?

Location in the network

AI compute can be located at various points in a network, for example centrally (i.e., on-premises at data centres), or at the edge (i.e., through connected mobile edge devices). Location usually determines the proximity of the user to the data and compute. AI compute at the edge is close to the user but might be less efficient as devices are decentralised and often not specialised for AI. In contrast, central compute locations tend to be further from users but offer greater capacity and capabilities than edge devices. The following questions could be considered:

- Where is AI compute capacity located in country networks – primarily in data centres, at the edge, or a mix of both?
- How have trends in where AI compute is located changed over time and what does this reveal about changing demand for AI compute?
- How can central AI compute be measured in the network? Is it possible to measure all on-premise data centres with AI workloads in a country?
- How can AI compute capacity at the edge be measured, given its decentralised nature and the multiplicity of devices such as personal mobile devices, laptop computers, and IoT devices? How can this support efficiency objectives?

Ownership

Compute for AI can be privately or publicly owned, for instance through commercial cloud providers offering compute services, and HPC resources located and owned at publicly funded and publicly accessible institutions such as universities or academic centres. In recent years, governments have explored investments in public compute resources, for example through public research cloud initiatives (Zhang et al., 2022[12]) and by bringing together expert groups, such as the United States National AI Research Resource (NAIRR) task force, to inform policy. The following questions could be considered:

- How does AI compute ownership differ between private and public (i.e., academic and research) provided resources?

- Who are the primary providers of AI compute in the country and what are the ownership models (e.g., service providers, renting capacity, building infrastructure, hardware providers, etc.)?
- How does access to AI compute vary depending on ownership? Are there barriers associated with each, such as cost, expertise, location etc.?

Supply chains

The AI compute supply chain comprises stages such as the extraction of natural resources, hardware manufacturing and processing (i.e., in semiconductor facilities), transportation, hardware assembly for availability in the cloud, and more. Countries increasingly focus on securing supply chains to avoid production bottlenecks and shortages. The following questions could be considered:

- How do the parts of AI compute supply chains relate to a particular national context?
- How robust are these supply chains? Is there sufficient contingency in global or domestic markets to source resources in the event of global shocks?

Sustainability (efficiency, environmental impacts)

Training and using large AI systems require significant compute resources, leading to environmental impacts: energy and water use, carbon emissions, e-waste, and natural resource extraction like rare mineral mining. This is a concern especially considering the rapidly growing compute needs of AI systems. Several good practices for sustainable AI exist, such as using pre-trained models where appropriate and powering data centres with renewable energy. Efficiency gains should be explored for compute hardware and software, including algorithms. For example, researchers at the Massachusetts Institute of Technology (MIT) and start-up MosaicML are training neural networks up to seven times faster by configuring AI algorithms to learn more efficiently. This topic is further explored in a 2022 report informed by the Expert Group (OECD, 2022[14]). The following questions could be considered:

- How resource-intensive is existing AI compute capacity (e.g., energy, water, carbon emissions etc.)? What portion of natural resource use is attributed to AI specifically, compared to ICTs in general?
- Can the national energy grid support future AI compute needs in a sustainable way? Have policies been considered that set design standards to minimise energy use and environmental impacts?
- How can existing lifecycle impact assessments and standards be leveraged to measure environmental impacts of AI compute and applications?
- What efficiency gains could be achieved by applying infrastructure and hardware best-practices and changes at the AI model level?

Additional considerations

Depending on their national AI context and level of technology adoption, policy makers might consider additional factors, such as the type of AI model, AI applications, stage of the AI system lifecycle, and access to data. These considerations can be integrated into AI compute plans to fit varying national contexts.

AI model type

AI can be enabled by different methods, such as symbolic AI methods or ML – the most popular method today for creating AI which includes sub-methods such as deep learning. Hybrid options are also possible.

- How do different types of AI methods (e.g., machine learning, symbolic AI, and hybrid AI) impact AI compute needs?

AI applications

AI can require specialised compute depending on its application, for instance whether an AI is being used for NLP, computer vision, robotics, the generation of recommendations, optimisation, or other applications.

- How do AI applications impact AI compute needs?

AI system lifecycle stage

The AI system lifecycle encompasses the following phases: (1) plan and design; (2) collect and process data; (3) build and use the model; (4) verify and validate the model; (5) deploy; and (6) operate and monitor the system (OECD, 2022[5]). Compute needs change with the phases of the AI system lifecycle, notably depending on whether an AI model is being trained or used for inferencing.

- How do domestic and sectoral AI compute needs differ along the AI system lifecycle (e.g., for training or inferencing)?

Data access and processing

Access to data for AI training and use, and the compute capacity needed to process and clean data for AI model training are key considerations. Along with algorithms and compute, data is an enabler of AI. Data localisation rules might require data to be physically stored in-country, potentially creating challenges for its use to train and deploy AI models. While the present paper focuses on AI compute, ensuring access to sufficient data and safeguards for its responsible use are essential, as articulated in the OECD AI Principles, and is the subject of separate OECD work (OECD, 2019[3]; OECD, 2015[23]).

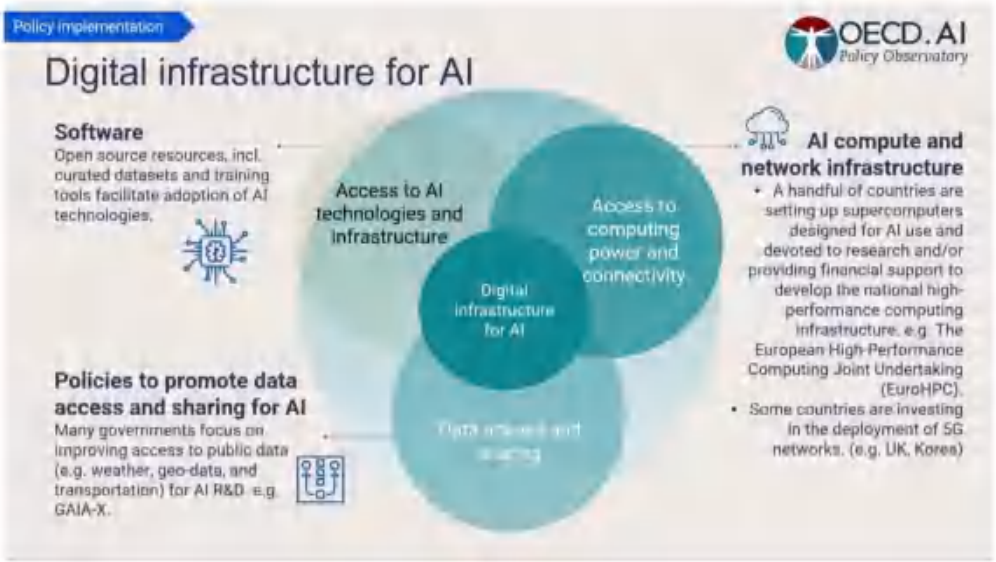
- What is the impact of sovereignty considerations and data localisation requirements on whether data is physically stored in-country (or in-region in the case of the European Union)?
- What challenges exist related to the compute needs of data for AI training and deployment?
- How much compute capacity is required to clean and prepare data for AI training or deployment?

5 AI compute in national policy initiatives

Countries and regions take varying approaches to providing the digital infrastructure and access required for the development and use of artificial intelligence (AI) (Figure 1). Different national goals for AI lead to different investment strategies. From building domestic infrastructure, to investments in the cloud, countries consider compute infrastructure investments on a case-by-case basis corresponding to national objectives. National AI initiatives related to computing resources often focus on general research and science infrastructure rather than AI specifically. While several countries have broader national HPC or cloud resources initiatives, few national AI plans have specifically targeted initiatives for assessing national AI compute capacity and needs.

This section is informed the OECD AI Policy Observatory, which includes a database of over 800 AI policy initiatives from more than 69 countries and territories, and the European Union. The database collects qualitative and quantitative data on national trends in AI policy. It includes a taxonomy for classifying policy initiatives according to four themes: (1) governance; (2) financial support; (3) AI enablers and other incentives; and (4) guidance and regulation (OECD, 2022[21]). In 2021, a new category was added called "AI computing and research infrastructure"' to collect information on related national AI initiatives. AI compute data and analysis on the OECD AI Policy Observatory will expand as awareness grows around including AI compute considerations in national plans.

Figure 1. Digital infrastructure for AI



Note: This stylised figure from the OECD 2021 report on the State of Implementation of the OECD AI Principles identifies a selection of AI policy instruments used by countries to implement OECD AI Principle 2.2 on fostering a digital ecosystem for AI.
Source: (OECD, 2021[25])

5.1. High-performance computing initiatives

HPC initiatives can be found across many countries and regions. Their focus is often on supporting a range of scientific and mathematical applications in addition to AI-specific initiatives.

Canada's Pan-Canadian Artificial Intelligence Strategy, launched in 2017 and renewed in 2021, leverages a national network of AI research institutes and supports the acquisition of HPC capacity dedicated for AI researchers. Canada's Advanced Research Computing Expansion Program launched in 2019 provides an increase in general national HPC capacity through Canada's supercomputing platform through the University of Victoria, Simon Fraser University, University of Waterloo, University of Toronto and McGill University, and coordinated by the Digital Research Alliance of Canada. In 2020, the first Canadian Digital Research Infrastructure Needs Assessment was launched to identify and address future digital research infrastructure and service needs (Digital Research Alliance of Canada, 2020[26]).

Chile established a National Laboratory for High Performance Computing to consolidate a national facility for HPC to help meet the national demand for computing resources from the scientific community. It offers services for both basic and applied research, with an emphasis on industrial applications (National Laboratory for High Performance Computing - Chile, n.d.[27]).

Colombia's Ministry of Information Technology and Communications is establishing the Colombian Supercomputing Network (Analític4), accessible to public- and private-sector actors (Ministerio de Tecnologías de la Información y las Comunicaciones, 2021[28]).

In **France**, the Grand Équipement National De Calcul Intensif created in 2007 is charged with providing HPC storage and services to researchers, academia, and industry for large-scale mathematical simulations, data processing, science, and AI applications (GENCI, n.d.[29]).

Germany's AI Strategy and HPC-Programme of the Federal Ministry for Education and Research aim to build national compute capacity through several national supercomputing centres, such as expanding the Gauss Centre for Supercomputing to exascale capability (Federal Government of Germany, 2020[30]).

In **Japan**, the RIKEN Centre for Computational Science and Fujitsu launched a top-ranked supercomputer named Fugaku in 2020. The National Institute of Advanced Industrial Science and Technology (AIST) develops and operates open AI computing infrastructure, including an initiative named AI Bridging Cloud Infrastructure to accelerate collaborative AI R&D between industry, academia, and the government (OECD, 2021[31]).

Korea announced their National High-Performance Computing Innovation Strategy for the Quantum Jump of the Fourth Industrial Revolution in May 2021. It consists of a 10-year medium- to long-term plan to close the gap with leading countries and create growth opportunities in line with domestic and global technology shifts, such as the transition to exascale computing, strengthening technological security, and increasing domestic demand.

Slovenia's National Supercomputing Network (SLING) provides national capacity for HPC compute to university and industry researchers, providing access to international and domestic cluster-based storage and compute capabilities (SLING, n.d.[31]). The 2020-25 Slovenian national AI strategy recognises compute infrastructure, including HPC and storage, as key. In 2017, the State Centre for Data Management and Storage was created to provide government offices with access to a State Cloud (Republic of Slovenia, 2021[32]).

In **Spain**, the Barcelona Supercomputing Centre established in 2004 provides HPC services to scientists and industry, with a pre-exascale system to be operational in 2023 and is a European leader in computer architectures research and HPC for AI applications (Barcelona Supercomputing Centre, 2022[33]).

In October 2020, the **United Kingdom** announced the launch of its most powerful supercomputer for use by healthcare researchers to tackle pressing medical challenges (OECD, 2021[25]). In 2022, the United

Kingdom launched a review of its digital research infrastructure needs to support the development and use of AI, examining the provision of compute, data access, and talent, which will inform its ongoing national AI strategy (The Alan Turing Institute, 2022[34]).

In 2022, the **United States** Department of Energy launched the Frontier supercomputer as one of the world's most powerful HPCs for AI applications (US Department of Energy, 2019[35]). The National Science Foundation (NSF) invests significantly in next-generation AI R&D supercomputers, such as Frontera, deployed in June 2019 (National Science Foundation, 2019[36]), and provides programs for access to AI compute through the National AI Research Institutes (National Science Foundation, 2022[37]). The National Aeronautics and Space Administration (NASA) has a high-end computing programme and is augmenting its Pleiades supercomputer with new nodes specifically designed for machine-learning (ML) AI workloads (OECD, 2021[31]). The United States National AI Initiative Act of 2020 plans to make world-class computing resources and datasets available to researchers across the country through the forthcoming United States National AI Research Resource (NAIRR).

India's Centre of Excellence in Artificial Intelligence is developing the National Artificial Intelligence Resource Portal, which will offer a web-based system to search and browse AI resources, including training and a cloud-based compute platform (Centre for Excellence in Artificial Intelligence, 2022[38]).

In **Serbia**, plans were announced to establish a National Platform for AI, including high-performance supercomputing capacity, accessible available upon request to certain institutions and the private sector (The Government of the Republic of Serbia, 2020[39]).

In **Thailand**, the National Science and Technology Development Agency (NSTDA) created Thailand's Supercomputer Centre (ThaiSC) in 2019 to provide national-scale supercomputing resources for R&D located in the Thailand Science Park (ThaiSC, 2022[35]).

In **Europe**, the European High-Performance Computing Joint Undertaking (EuroHPC) was established in 2018 to share computing resources and coordinate efforts among EU countries and partners, with a 2021-27 budget of EUR 7 billion (EuroHPC, 2022[22]). It aims to develop peta and pre-exa-scale supercomputing capacities, and data infrastructure to support European scientific and industrial research and innovation for scientific, industrial, and public users, including for AI (OECD, 2021[12]; EuroHPC, 2022[11]). Launched in 2021, the EU-ASEAN High-Performance Computing Virtual School hosted by ThaiSC brings together experts, students, and researchers from Europe and ASEAN member states to share best practices and learn HPC design and programming skills (The ASEAN Secretariat, 2021[23]).

5.2. Cloud-based services

Initiatives exist to address other important enablers for AI compute, including data processing, broadband networks, and cloud-based services. The OECD Going Digital Toolkit defines cloud computing as "ICT services over the Internet to access servers, storage, network components and software applications" (OECD Going Digital Toolkit, 2021[43]).

In 2019, **France** and **Germany** launched GAIA-X, an EU cloud-based initiative that aims to establish an interoperable data exchange through which business and research partners can share data and access services at scale, including for AI (Gaia-X, n.d.[44]).

Cloud computing and connectivity initiatives can be found across **Europe** for a variety of uses. Since 2016, the European Commission has been developing a blueprint for cloud-based services and data infrastructure, including the European Data Infrastructure and the European Open Science Cloud, which will deploy high-bandwidth networks, large scale storage, and supercomputer capacity for academic and industry partners (data.europa.eu, 2016[45]).

5.3. Supply chain initiatives

In addition to national and regional investments in HPC and cloud service capabilities, initiatives are increasingly being launched to secure upstream manufacturing of components for AI compute, such as securing semiconductor supply chains:

- In 2020, **Korea** launched its AI Semiconductor Industry Development Strategy, a USD 1 billion cross-ministerial project. As part of the Digital Korea Strategy launched in 2022, **Korea** is planning a K-Cloud Project, which operates a cloud data centre established with domestically developed semiconductors to promote AI infrastructure and services (Ministry of Science and ICT, n.d.[46]).
- Aligned with other European initiatives, such as the European Chips Act (below), **Spain** approved a strategic plan of more than EUR 12 billion to develop the design and production capacities of the Spanish microelectronics and semiconductor industry, covering the value chain from design to chip manufacturing.
- The **United States** established the Creating Helpful Incentives to Produce Semiconductors (CHIPS) for America Act, which offers USD 52 billion for semiconductor manufacturing, supply chain and R&D investments (Congress.Gov, 2020[47]).
- The **European Union** announced the European Chips Act to incentivise over EUR 15 billion in public and private sector investments (European Commission, 2019[48]).

6 Gap analysis and preliminary findings

Making informed and evidence-based decisions to plan national compute capacity for the fast-changing needs of AI systems can be challenging. A suite of indicators and proxies will be needed to measure national AI compute capacity and preparedness to meet AI goals. This section identifies gaps in existing measurement tools and discusses preliminary findings.

6.1. AI policy initiatives need to take AI compute capacity into account

At present, national AI policy initiatives do not include detailed measures of AI compute capacity and corresponding national needs, focusing instead on general-purpose compute. As such, measuring and planning for the AI compute needed to realise national AI plans is challenging and relies on high-level strategic goals articulated in national AI strategies. Translating the AI ambitions contained in such plans into more concrete considerations – such as reviewing current national compute capacity and the AI compute needs of public and private sector actors – would enable more efficient and targeted planning of AI compute investments. Consideration should also be given to measuring whether national AI compute is owned domestically or rented from providers abroad, such as through cloud services. Based on national needs and security priorities, attention to domestically owned compute capacity could be warranted.

6.2. National and regional data collection and measurement standards need to expand

Data collection should be expanded to measure current national AI compute capacity and needs, particularly at national and regional levels. This could include measuring existing private and public HPC clusters (including the number of data centres used to support AI workloads), which could be aggregated to provide insights on a national level. Data collection should follow measurement standards, using consistent terminology, indicators, and metrics to allow for comparability across jurisdictions. Collaboration between private-sector actors, governments, national statistical offices, academia, and the OECD could support such data-collection efforts.

The AI compute indicators under discussion proposed in Annex C could provide insights into national compute capacity and needs. However, data associated with many of these indicators might not be publicly available nor aggregated at the national level. For example, insights into private cloud computing capacities and the number of AI compute hardware customers might be deemed commercially sensitive. To build on existing data-collection efforts, analysis of the activities of national statistical offices related to measuring AI compute could be explored.

6.3. Policy makers need insights into the compute demands of AI systems

Further insights are needed into the compute demands for both the training and inferencing stages of an AI system's lifecycle. Most data on AI compute focuses on training. While compute for AI training is critical and requires significant computing resources within limited timeframes, inferencing can also use significant

AI compute resources over an AI system's lifecycle (Patterson et al., 2021[17]; Patterson, 2022[18]). Further focus is needed on the compute requirements of AI systems during the various lifecycle stages – notably data processing, development, deployment, and operation – along with analysis and forecasting of current and future AI compute demands, so that policy makers and others can plan accordingly.

6.4. AI-specific measurements should be differentiated from general-purpose compute

Identifying the differences between AI compute and general-purpose compute is challenging. Untangling these measures would allow countries to quantify their existing AI compute capacity and allow for more strategic coordination with other plans, for instance regarding investments into compute infrastructure for advanced science and mathematical modelling. This could allow countries to leverage synergies between AI-specific compute and general-purpose compute.

6.5. Workers need access to AI compute related skills and training

AI compute hardware alone is not sufficient to enable the development and deployment of AI. Users, such as researchers and developers, need to be able to adequately access AI compute and related support services to efficiently and effectively utilise HPC clusters. Very specific skills are often needed, such as from engineers and those with experience using specialised hardware for AI. Perspectives from diverse disciplines and backgrounds are also critical to close compute divides between developed and emerging economies, and the public, academic, and private sectors. Research is needed into the supply of and demand for AI compute skills, training, and workforce composition to understand what investments might support the full effective use of national AI compute capacity.

6.6. AI compute supply chains and inputs need to be mapped and analysed

As countries scale up AI compute capacity according to national needs, demand for various inputs along AI compute supply chains could increase. This could reveal bottlenecks and resource constraints, as illustrated by challenges surrounding the semiconductor industry. AI compute supply chains and inputs require further mapping and analysis so governments can build contingency and resilience plans.

7 Conclusion

AI is a general-purpose technology impacting nearly every facet of the global economy, prompting governments to formulate and publish national AI strategies. The successful implementation of national AI strategies could become one of the factors defining a country's ability to deliver innovation, productivity gains, and long-term growth. Governments are allocating budgets and investing public funds to support the implementation of such AI strategies and programs.

However, many countries have developed AI plans without a full assessment of whether they have sufficient domestic AI compute capacity to realise these goals. Concerns are growing about reinforcing divides between those who have the resources to create and use complex AI models to generate competitive advantage and productivity gains, and those who do not. Without data on national compute capacity and the needs of AI ecosystems, decisionmakers might not be able to effectively implement and leverage strategic national AI investments and plans for economic growth and competitiveness.

Understanding of AI compute and its relationship to the diffusion of AI across OECD and partner economies can improve the implementation of national AI strategies, and guide future policymaking and investments. Countries should consider systematically taking stock of existing national compute capacity and reviewing the current and emerging needs of their AI ecosystem. National AI compute plans based on common definitions, standards, and data collection can equip governments and policy makers to make informed decisions in a fast-changing global digital economy, and close compute divides around the world.

Notes

[1] Calculations by OpenAI estimate that "since 2012, the amount of compute used in the largest AI training runs has been increasing exponentially with a 3.4-month doubling time (by comparison, Moore's Law had a 2-year doubling period). Since 2012, this metric has grown by more than 300,000x (a 2-year doubling period would yield only a 7x increase)." For more details see (OpenAI, 2018[1]).

[2] In November 2009, the leading supercomputer ranked as #1 in the Top500 (called Jaguar) demonstrated a performance of 1.75 peta floating-point operations per second (PFLOPS). In November 2022, the leading supercomputer demonstrated a performance of 1 102 PFLOPS according to the Top500 (called Frontier). Therefore, growth by a factor of approximately 630 between these two supercomputers can be calculated (i.e., 1 102 PFLOPS / 1.75 PFLOPS). For further detail, please visit: https://www.top500.org/statistics/perfdevel/

[3] A large cloud compute provider does not wish to be attributed by name due to commercial confidentiality concerns.

References

Advanced Research Computing (ARC) (2022), *What is High Performance Computing?*, US. Geological Survey, https://www.usgs.gov/advanced-research-computing/what-high-performance-computing (accessed on 22 April 2022). [52]

Ahmed, N. and M. Wahed (2020), "The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research", https://arxiv.org/abs/2010.15581. [4]

Barcelona Supercomputing Centre (2022), *What we do*, https://www.bsc.es/discover-bsc/the-centre/what-we-do. [33]

Batra, G. et al. (2018), *Artificial-intelligence hardware: New opportunities for semiconductor companies*, McKinsey & Company, https://www.mckinsey.com/~/media/McKinsey/Industries/Semiconductors/Our%20Insights/Artificial%20intelligence%20hardware%20New%20opportunities%20for%20semiconductor%20companies/Artificial-intelligence-hardware.ashx (accessed on 25 April 2022). [60]

Bengio, Y., A. Courville and I. Goodfellow (2016), *Deep Learning*, MIT Press, https://mitpress.mit.edu/books/deep-learning. [16]

Centre for Excellence in Artificial Intelligence (2022), *National Artificial Intelligence Resource Portal (NAIRP)*, http://www.ai.iitkgp.ac.in/Research/. [38]

Cisco (2020), *Cisco Annual Internet Report (2018–2023) White Paper*, https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html. [72]

Cisco (2018), *Cisco Global Cloud Index: Forecast and Methodology, 2016–2021*, Cisco. [71]

Cisco (n.d.), *What Is a Data Center?*, https://www.cisco.com/c/en/us/solutions/data-center-virtualization/what-is-a-data-center.html (accessed on 25 April 2022). [63]

Congress.Gov (2020), *H.R.7178 - CHIPS for America Act*, https://www.congress.gov/bill/116th-congress/house-bill/7178#:~:text=This%20bill%20establishes%20investments%20and,manufacturing%20facility%20investment%20through%202026. [47]

data.europa.eu (2016), *Europe's blueprint for Cloud services*, https://data.europa.eu/en/news/europes-blueprint-cloud-services. [45]

Dell Technologies (2021), *Global Data Protection Index 2021*, Dell Technologies. [68]

Digital Research Alliance of Canada (2020), *Canadian Digital Research Infrastructure Needs Assessment*, https://alliancecan.ca/en/initiatives/canadian-digital-research-infrastructure-needs-assessment. [26]

EuroHPC (2022), *Discover EuroHPC JU*, https://eurohpc-ju.europa.eu/discover-eurohpc-ju. [41]

European Commission (2019), *European Chips Act*, https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-chips-act_en. [48]

Eurostat (2021), *Cloud computing - statistics on the use by enterprises*, https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Cloud_computing_-_statistics_on_the_use_by_enterprises#Use_of_cloud_computing:_highlights. [9]

Federal Government of Germany (2020), *Artificial Intelligence Strategy of the German Federal Government*, https://www.ki-strategie-deutschland.de/files/downloads/Fortschreibung_KI-Strategie_engl.pdf (accessed on 25 April 2022). [30]

Gaia-X (n.d.), *What is Gaia-X?*, https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html, https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html (accessed on 2022). [44]

Ganguli, D. et al. (2022), "Predictability and Surprise in Large Generative Models", https://arxiv.org/abs/2202.07785. [13]

Gartner (n.d.), *Gartner Glossary: CPU (Central Processing Unit)*, https://www.gartner.com/en/information-technology/glossary/cpu-central-processing-unit. [61]

Gartner (n.d.), *Gartner Glossary: Data Center*, https://www.gartner.com/en/information-technology/glossary/data-center (accessed on 25 April 2022). [64]

GENCI (n.d.), *GENCI*, https://www.genci.fr/en. [29]

Gill, B. and D. Smith (2018), *The Edge Completes the Cloud: A Gartner Trend Insight Report*, Gartner, https://emtemp.gcom.cloud/ngw/globalassets/en/doc/documents/3889058-the-edge-completes-the-cloud-a-gartner-trend-insight-report.pdf (accessed on 25 April 2022). [59]

Heim, L. (2021), *What is Compute?*, AI Alignment Forum, https://www.alignmentforum.org/posts/uYXAv6Audr2y4ytJe/what-is-compute-transformative-ai-and-compute-1-4 (accessed on 25 April 2022). [49]

He, X., M. He and Z. Han (2018), *A Survey of Network Topology of Data Center*, 2018 IEEE 4th International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing, (HPSC) and IEEE International Conference on Intelligent Data and Security (IDS), https://doi.org/10.1109/BDS/HPSC/IDS18.2018.00021. [65]

Hoffmann, J. et al. (2022), "Training Compute-Optimal Large Language Models", https://arxiv.org/abs/2203.15556. [7]

IBM (n.d.), *What is Edge Computing*, https://www.ibm.com/cloud/what-is-edge-computing (accessed on 25 April 2022). [58]

IEA (2021), *Data Centres and Data Transmission Networks*, IEA, https://www.iea.org/reports/data-centres-and-data-transmission-networks. [69]

IEA (2017), *Digitalisation and Energy*, IEA, https://www.iea.org/reports/digitalisation-and-energy. [70]

International Labour Organization (2016), *International Standard Classification of Occupations*, https://www.ilo.org/public/english/bureau/stat/isco/isco08/. [20]

International Telecommunication Union (2019), *Cloud computing: From paradigm to operation*, https://www.itu.int/en/publications/Documents/tsb/2020-Cloud-computing-From-paradigm-to-operation/files/downloads/Cloud-computing-20-00081E.pdf (accessed on 25 April 2022). [56]

ISO and IEC (2021), *ISO/IEC TR 24030:2021 Information technology — Artificial intelligence (AI) — Use cases*, https://www.iso.org/standard/77610.html. [76]

Kaplan, J. et al. (2020), "Scaling Laws for Neural Language Models", https://arxiv.org/abs/2001.08361. [6]

Ker, D. (2021), "Measuring cloud services use by businesses", *OECD Digital Economy Papers,* No. 304, https://doi.org/10.1787/71a0eb69-en. [73]

Khan, S. and A. Mann (2020), *AI Chips: What They Are and Why They Matter*, https://cset.georgetown.edu/wp-content/uploads/AI-Chips%E2%80%94What-They-Are-and-Why-They-Matter.pdf. [12]

Khan, S., A. Mann and D. Peterson (2021), *The Semiconductor Supply Chain: Assessing National Competititveness*, Center for Security and Emerging Technology, https://cset.georgetown.edu/wp-content/uploads/The-Semiconductor-Supply-Chain-Issue-Brief.pdf. [10]

Komprise (2022), *AI Compute*, Data Management Glossary, https://www.komprise.com/glossary_terms/ai-compute/ (accessed on 25 April 2022). [51]

Law Insider (2022), *Compute Capacity definition*, https://www.lawinsider.com/dictionary/compute-capacity (accessed on 25 April 2022). [50]

Ministerio de Tecnologías de la Información y las Comunicaciones (2021), *Analític4, la red colombiana de supercomputación que acompañará a la industria y al Gobierno para generar soluciones basadas en análisis de datos*, https://mintic.gov.co/portal/inicio/Sala-de-prensa/Noticias/194433:Analitic4-la-red-colombiana-de-supercomputacion-que-acompanara-a-la-industria-y-al-Gobierno-para-generar-soluciones-basadas-en-analisis-de-datos. [28]

Ministry of Science and ICT (n.d.), *Korea to Come up with the Roadmap of Digital ROK, Realizing the New York Initiative*, https://www.msit.go.kr/eng/bbs/view.do?sCode=eng&mId=4&mPid=2&bbsSeqNo=42&nttSeqNo=742. [46]

MIT Technology Review and Infosys Cobalt (2022), *The Global Cloud Ecosystem Index 2022*, MIT Technology Review and Infosys Cobalt, https://wp.technologyreview.com/wp-content/uploads/2022/04/MITTR-INFOSYS-Cloud_Reort_FNL.pdf?fbclid=IwAR1LgGMak430HwmgV6C0qmXM_J6iTe3tg2eO9EZpa8rdIGhL7LuBXTu19sE#:~:text=The%20Global%20Cloud%20Ecosystem%20Index%202022%20is%20a%20snapshot%20of,to%20promote. [75]

MLCommons (2022), *Machine learning innovation to benefit everyone.*. [74]

National Center for High Performance Computing (2022), *High Performance Computing (HPC)*, Istanbul Technical University National Center for High Performance Computing, https://en.uhem.itu.edu.tr/ybh.html (accessed on 25 April 2022). [53]

National Institute for Research in Digital Science and Technology (2021), *The essentials on: high performance computing*, https://www.inria.fr/en/essential-high-performance-computing-hpc. [54]

National Institute of Standards and Technology (2011), *The NIST Definition of Cloud Computing*, https://csrc.nist.gov/publications/detail/sp/800-145/final#:~:text=Cloud%20computing%20is%20a%20model,effort%20or%20service%20provider%20interaction (accessed on 25 April 2022). [77]

National Institute of Standards and Technology (2011), *The NIST Definition of Cloud Computing*, https://www.govinfo.gov/app/details/GOVPUB-C13-74cdc274b1109a7e1ead7185dfec2ada#:~:text=Cloud%20computing%20is%20a%20model,effort%20or%20service%20provider%20interaction. [57]

National Laboratory for High Performance Computing - Chile (n.d.), *Laboratorio Nacional de Computación de Alto Rendimiento*, https://www.nlhpc.cl/. [27]

National Science Foundation (2022), *National Artificial Intelligence Research Institutes*, https://beta.nsf.gov/funding/opportunities/national-artificial-intelligence-research-institutes. [37]

National Science Foundation (2019), *NSF-funded leadership-class computing center boosts U.S. science with largest academic supercomputer in the world*, https://www.nsf.gov/news/news_summ.jsp?cntn_id=299134. [36]

OECD (2022), "Measuring the environmental impacts of AI compute and applications: The AI footprint", *OECD Digital Economy Papers,* No. 341, https://doi.org/10.1787/7babf571-en. [14]

OECD (2022), *OECD AI Policy Observatory*, https://oecd.ai/en/. [24]

OECD (2022), "OECD Framework for the Classification of AI systems", *OECD Digital Economy Papers,* No. 323, https://doi.org/10.1787/cb6d9eca-en. [15]

OECD (2021), "State of Implementation of the OECD AI Principles: Insights from National AI Policies", *OECD Digital Economy Papers,* No. 311, https://doi.org/10.1787/1cd40c44-en. [25]

OECD (2019), "Measuring distortions in international markets: The semiconductor value chain", *OECD Trade Policy Papers*, No. 234, OECD Publishing, Paris, https://doi.org/10.1787/8fe4491d-en. [11]

OECD (2019), *Recommendation of the Council on Artificial Intelligence*, OECD, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (accessed on 12 April 2022). [3]

OECD (2015), *Data-Driven Innovation: Big Data for Growth and Well-Being*, OECD Publishing, https://doi.org/10.1787/9789264229358-en. [23]

OECD (2014), "Cloud Computing: The Concept, Impacts and the Role of Government Policy", *OECD Digital Economy Papers,* No. 240, https://doi.org/10.1787/5jxzf4lcc7f5-en. [55]

OECD Going Digital Toolkit (2021), *Share of businesses purchasing cloud services*, https://goingdigital.oecd.org/indicator/25. [43]

Online Browsing Platform (1993), *Information technology — Vocabulary — Part 1: Fundamental terms*, https://www.iso.org/obp/ui/#iso:std:iso-iec:2382:-1:ed-3:v1:en (accessed on 25 April 2022). [62]

OpenAI (2018), *AI and Compute*, https://openai.com/blog/ai-and-compute/. [1]

Oxford Insights (2021), *Government AI Readiness Index 2021*, Oxford Insights. [67]

Patterson, D. (2022), "The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink", *TechRxiv*, https://www.techrxiv.org/articles/preprint/The_Carbon_Footprint_of_Machine_Learning_Training_Will_Plateau_Then_Shrink/19139645/1 (accessed on 21 April 2022). [18]

Patterson, D. et al. (2021), "Carbon Emissions and Large Neural Network Training", https://arxiv.org/abs/2104.10350. [17]

Republic of Slovenia (2021), *Slovenian State Cloud DRO*, https://nio.gov.si/nio/asset/drzavni+racunalniski+oblak+dro?lang=en. [32]

Samek, L., M. Squicciarini and E. Cammeraat (2021), "The human capital behind AI: Jobs and skills demand from online job postings", *OECD Science, Technology, and Industry Policy Papers,* No. 120, https://doi.org/10.1787/2e278150-en. [21]

Sevilla, J. et al. (2022), "Compute Trends Across Three Eras of Machine Learning", https://arxiv.org/abs/2202.05924 (accessed on 22 February 2023). [2]

SLING (n.d.), *Slovenian national supercomputing network*, https://www.sling.si/sling/en/. [31]

ThaiSC (2022), *NSTDA Supercomputer Center*, https://thaisc.io/en/mainpage/. [40]

The Alan Turing Institute (2022), *UK AI Research Infrastructure Requirements Review*, https://www.turing.ac.uk/work-turing/uk-ai-research-infrastructure-requirements-review. [34]

The ASEAN Secretariat (2021), *EU, ASEAN kick off first high-performance computing school*, https://asean.org/eu-asean-kick-off-first-high-performance-computing-school/. [42]

The Government of the Republic of Serbia (2020), *Strategy for the Development of Artificial Intelligence in the Republic of Serbia for the period 2020-2025*, [39]

https://www.srbija.gov.rs/tekst/en/149169/strategy-for-the-development-of-artificial-intelligence-in-the-republic-of-serbia-for-the-period-2020-2025.php.

Thompson, N. et al. (2020), "The Computational Limits of Deep Learning", https://arxiv.org/abs/2007.05558. [8]

Top500 (2022), *November 2022 List*, https://www.top500.org/lists/top500/2022/11/ (accessed on January 2023). [5]

Tortoise Media (2022), *The Global AI Index*, https://www.tortoisemedia.com/intelligence/global-ai/. [66]

United States Census Bureau (2022), *North American Industry Classification System*, https://www.census.gov/naics/?input=54&chart=2022&details=541511. [19]

US Department of Energy (2019), *U.S. Department of Energy and Cray to Deliver Record-Setting Frontier Supercomputer at ORNL*, https://www.energy.gov/articles/us-department-energy-and-cray-deliver-record-setting-frontier-supercomputer-ornl. [35]

Zhang, D. et al. (2022), *The AI Index 2022 Annual Report*, AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University, https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf. [22]

Annex A. Examples of existing keyword definitions

The table below lists examples of keyword definitions from a variety of sources. This list does not constitute an endorsement by the OECD.AI Expert Group on AI Compute and Climate nor by the OECD. It is provided to illustrate the ways terms are defined by various actors and organisations. Existing keyword definitions were consulted in the Expert Group's discussion of a proposed definition of AI compute, and analysis will continue to refine and iterate on this work.

Terminology	Definition
Compute	• "Compute is the manipulation of information or any type of calculation — involving arithmetical and non-arithmetical steps. It can be seen as happening within a closed system: a computer. Examples of such physical systems include digital computers, analogue computers, mechanical computers, quantum computers, or wetware computers (your brain)." – (Heim, 2021[49]) • "Compute capacity means the physical or logical allocation of storage or processing power." – (Law Insider, 2022[50]) • "The computing ability required for machines to learn from big data to experience, adjust to new inputs, and perform human-like tasks." – (Komprise, 2022[51])
High-performance computing (HPC)	• "High Performance Computing most generally refers to the practice of aggregating computing power in a way that delivers much higher performance than one could get out of a typical computer or workstation in order to solve large problems in science, engineering, or business." – (Advanced Research Computing (ARC), 2022[52]) • "High Performance Computing (HPC), in the simplest term, is defined as distributing a computing job to multiple processors instead of running it on a single processor sequentially for a long duration." – (National Center for High Performance Computing, 2022[53]) • "High performance computing, also known as HPC, is the ability to perform complex calculations and massive data processing at very high speed by combining the power of several thousand processors" – (National Institute for Research in Digital Science and Technology, 2021[54])

Cloud computing	• "Computing services based on a set of computing resources that can be accessed in a flexible, elastic, on-demand way with low management effort." – (OECD, 2014[55]) • "Paradigm for enabling network access to a scalable and elastic pool of shareable physical or virtual resources with self-service provisioning and administration on-demand. Examples of resources include servers, operating systems, networks, software, applications, and storage equipment." – (International Telecommunication Union, 2019[56]) • "Cloud computing is a model for enabling ubiquitous, convenient on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" – (National Institute of Standards and Technology, 2011[57])
Edge computing	• "Edge computing is a distributed computing framework that brings enterprise applications closer to data sources such as IoT devices or local edge servers. This proximity to data at its source can deliver strong business benefits, including faster insights, improved response times, and better bandwidth availability." – (IBM, n.d.[58]) • "Cloud computing and edge computing are complementary, rather than competitive or mutually exclusive. Organizations that use them together will benefit from the synergies of solutions that maximize the benefits of both centralized and decentralized models. Edge computing will take place at the absolute edge, and it will be leveraged anywhere in a distributed computing architecture that meets use case requirements for latency, bandwidth, data privacy and autonomy." – (Gill and Smith, 2018[59])
Processor	• "Compute performance relies on central processing units (CPUs) and accelerators—graphics-processing units (GPUs), field programmable gate arrays (FPGAs), and application-specific integrated circuits (ASICs)." – (Batra et al., 2018[60]) • "The component of a computer system that controls the interpretation and execution of instructions. The CPU of a PC consists of single microprocessor, while the CPU of a more powerful mainframe consists of multiple processing devices, and in some cases, hundreds of them. The term "processor" is often used to refer to a CPU." – (Gartner, n.d.[61]) • "In a computer, a functional unit that interprets and executes instructions. A processor consists of at least an instruction control unit and an arithmetic and logic unit." – (Online Browsing Platform, 1993[62])
Data centre	• "…a data centre is a physical facility that organizations use to house their critical applications and data. A data centre's design is based on a network of computing and storage resources that enable the delivery of shared applications and data. The key components of a data centre design include routers, switches, firewalls, storage systems, servers, and application-delivery controllers." – (Cisco, n.d.[63]) • "A data centre is the department in an enterprise that houses and maintains back-end IT systems and data stores – its mainframes, servers, and databases. In the day of large, centralized IT operations, this department and all the systems resided in one physical place, hence the name data centre." – (Gartner, n.d.[64]) • "Data centre is used as a network infrastructure for carrying, transmitting, storing, and processing big data, which plays an important role in the application of cloud computing, CDN distribution, etc." – (He, He and Han, 2018[65])

Annex B. Existing datasets, indicators, and proxies for AI compute

The table below outlines datasets, indicators, and proxies for AI compute available at the time of writing. This list is not exhaustive and constitutes a sample of possible datasets and measurement tools that could be leveraged in future work.

Source	Indicators	Relevance to AI compute measurement categories
Stanford HAI, AI Index Report (Zhang et al., 2022[22]) Covers changes in R&D, AI model performance and applications, the economy (jobs, investing, industry), education, ethical challenges, diversity, policy, and national strategies.	Publication and conference counts, GitHub stars, patents (e.g., by region/country, sector), performance of state-of-art models in major domains, hiring and labor demand, investment, technology adoption, courses, PhDs, faculty in industry, number of ethics principles, media articles on AI, gender/race diversity in AI, international coalitions, public investment, legislation.	People (skills, training, diversity) Innovation (R&D) Policy (law, regulation, strategy)
Top 500 (Top500, 2022[5]) A list of the fastest supercomputers released twice a year. It compares computers using standardised performance tests and collects information on the hardware stack used including hardware vendors and countries of origin. Data collected biannually since 1993.	Number of processing cores, number of operations per second, power usage; prevalence of specific hardware, software, and interconnects; prevalence of hardware vendors; Green500 compares the number of operations made per watt (a power-efficiency benchmark). *Note: The benchmarks provide information on the computers, their site, manufacturer, hardware, and software setups. Data exists for the best-performing 500 HPCs, many of which make the list several times. Contributions to the list are voluntary, posing methodological challenges.*	Availability (supply) Sustainability (efficiency, environmental impacts)

The Global AI Index (Tortoise Media, 2022[66]) Examines AI development along three axes – implementation, innovation, and investment – and benchmark and ranks 62 countries. The measurement framework is aligned with the OECD "Handbook on constructing composite indicators" and based around capacity (as opposed to current use) for high AI output now and in the future.	LinkedIn skills reporting, GitHub stars, Scopus publications, number of Google searches about AI, STEM graduates (including gender), government spending, research tax credits, AI policies, funding to AI companies (startups/unicorns). *Note: The infrastructure indicators are a combination of internet speeds, Top500 data, and supply chain exports (e.g., chips). Several indicators are labelled as proxies, which are aggregated to rank countries and not necessarily AI-specific. Some indicators use information sources available on the OECD.AI Policy Observatory (e.g., Scopus).*	Availability (supply) People (skills, training, diversity) Policy (law, regulation, strategy) Innovation (R&D) Security & sovereignty (supply chains)
Government AI Readiness Index (Oxford Insights, 2021[67]) Uses 42 indicators across 10 dimensions for 9 geographic regions to measure and rank the readiness of 160 countries to implement AI for public services to citizens. Analysis is conducted on three pillars: government, technology, and data and infrastructure.	National AI strategy, data protection and privacy policy, national ethics framework, trust in government websites, software spending, number of unicorns and startups, R&D spending, digital skills, number of AI research papers, telecoms infrastructure and bandwidth, open government data (data availability), open data policies, gender gap (data representativeness). *Note: The index focuses on government readiness to use AI, not public- or private-sector use in general. The infrastructure pillar focus on telecommunications infrastructure and networking speed (which are necessarily AI-specific measures) and Top500 data.*	Availability (supply) People (skills, training, diversity) Policy (Law, regulation, policy) Innovation (R&D)
Global Data Protection Index (Dell Technologies, 2021[68]) Data collected from interviews of 1 000 IT decisionmakers around the world to gauge cloud security protection strategies and comfort.	Survey results on cost of data leak instances, unscheduled downtime, comfort with current cloud security setup, percentages of public/private cloud(s), security vendors, and emerging cloud services (including AI as a Service). *Note: Data relates primarily to security dimensions around cloud uses, applications, and growth in concern and cost of cloud cyberattacks and data protection.*	Availability (cloud) Security
Report of Data Centres and Data Transmission Networks (IEA, 2021[69]) Describes growth of network use and data services, current energy consumption of data centers, data center use of renewable energy, efficient data transfer (software virtualisation), and changing networks (e.g., growth of mobile).	Cloud vendors' renewable and total energy consumption, internet traffic over time, data center and network (fixed-line and mobile) electricity use in watt-hours. *Note: Data is measured globally (not by country) and does not contain data specific to AI compute. However, data gives trends of cloud energy consumption, accessibility, and efficiencies.*	Availability (supply) Sustainability (efficiency, environmental impacts)

Digitization and Energy (IEA, 2017[70]) Considers impact of digitisation across various industries, including ICT. Provides a cybersecurity, privacy, and economic disruption risk assessment.	Data center, network power and energy use over time; network traffic per year; mobile broadband subscriptions; internet access; investment in digital infrastructure; network-enabled smart-home technology. *Note: This source considers the impact of emerging technology (including ML on different industries) and rebound effects. The scope is broader than compute alone (with no specific mentions of AI compute). Includes data from Cisco, which may serve as proxies for measuring the hardware infrastructure supporting AI compute.*	Availability (supply) Sustainability (efficiency, environmental impacts) Security
Cisco Global Cloud Index (Cisco, 2018[71]) Forecasts growth of cloud traffic, data center traffic, compute, and data center storage.	Number and country of hyperscalers, datacenter traffic (to edge, among datacenters, within datacenters) per year for traditional and cloud data centers, and workloads and compute instances in the cloud; public/private cloud growth, service model trends, applications, storage, APIs, secure servers, and speeds and latencies. *Note: This source offers data on model service trends, cybersecurity, growth of hyperscale data centers, and cloud compute for application-specific (but no AI-specific) workloads at regional (not national) scale.*	Availability (cloud) Security & Sovereignty (location, ownership)
Cisco Annual Internet Report (2018-2023) (Cisco, 2020[72]) Forecasts network and telecommunications growth up to 2023. Includes security assessment which reviews some of AI's impacts and GDPR adoption.	Billions of internet users, devices, connection styles, IoT and mobile-to-mobile connections, mobile network growth, GDPR adoption survey data. *Note: This source focuses on networking, forecasting for internet access, edge, and mobile growth by region. Information is not necessarily AI-specific.*	Use (demand) Security and Sovereignty (location)
Measuring Cloud Services Use by Business, OECD (Ker, 2021[73]) Defines public cloud and cloud services according to various product classification frameworks and computes prices paid for product categories (e.g., data processing and internet publication) to proxy cloud use by business.	USD spent by businesses in different product categories, from information services to news and internet publishing. *Note: This source details analysis of data from country supply-use tables. Many countries' supply-use tables data is sparse or absent in broad product categories like information services and in narrower categories for cloud-specific uses and applications. The report aims to complement results for ICT survey-based data with results gleaned from economic and business statistics.*	Use (cloud demand) Policy (law, regulation, strategy)

MLCommons (MLCommons, 2022[74]) Provides benchmarking, datasets, and best-practices for measuring machine learning training and inference through a community of over 50 founding members and affiliates from the private sector, academia, and non-profits globally.	Training performance, including HPC, data centre, edge, and mobile inference. *Note: This source provides performance data for training and inference of machine learning systems through benchmarking such as MLPerf, which includes a benchmark suite measuring how fast systems can train models to a target quality metric and how fast systems can process inputs and produce results using a trained model. Datasets and best-practices are also provided.*	Availability (supply) Innovation (R&D and measure of performance)
MIT Global Cloud Ecosystem Index (MIT Technology Review and Infosys Cobalt, 2022[75]) Ranks 76 countries and how their policy, skills, and technology affect the availability of cloud services. Pillars are infrastructure, ecosystem adoption, security and assurance, and talent and human affinity.	Number of datacenters, secure datacenters, and IP addresses in a country; digital service adoption, SaaS growth, and price of broadband services; prevalence of engineering and mathematics skills, and internet and digital literacy. *Note: The pillars of analysis give composite scores for each country. These pillars are weighted (e.g., infrastructure pillar accounts for 15% of each country's score) and averaged into a single score to rank countries' cloud ecosystems. The specific data and methodology behind these scores are not included.*	Availability (supply) People (skills, training, diversity) Innovation (R&D) Access (access rights) Security and sovereignty (location)
ISO/IEC TR 24030:2021 (ISO and IEC, 2021[76]) Provides a selection of submitted AI use cases, focusing on trustworthiness (e.g., fairness and bias). Source behind paywall.	Distribution of submitted use cases (broken down by application, status of development, AI domain task (e.g., optimisation, NLP). *Note: This source focuses on the purpose of AI, including applications and trustworthiness, deployment models for AI (location in network and access), and AI use (model and application).*	Access (access rights) Security and sovereignty (location)

Annex C. Indicators under discussion

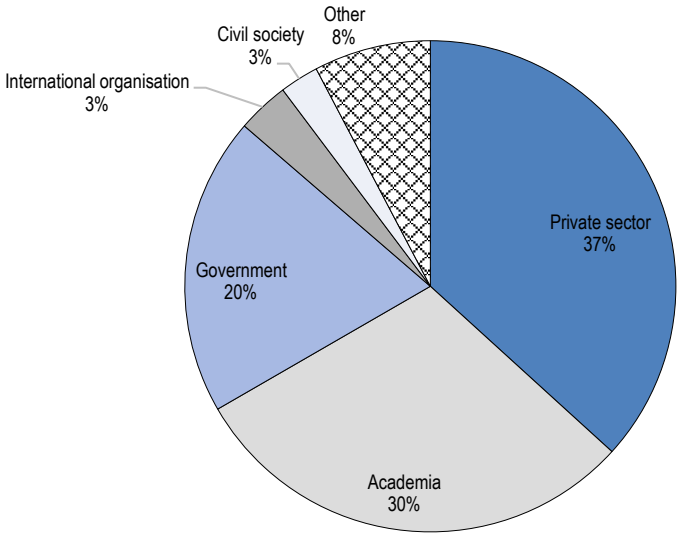
Informed by the considerations discussed for building a national AI compute plan, this list provides possible indicators and tools identified as corresponding to the proposed framework for measuring AI compute capacity availability and demand. Some of these metrics, indicators, and proxies already exist and are used within technical AI communities, whereas others are proposed for discussion and development. This list is under discussion by the OECD.AI Expert Group on AI Compute and Climate and will continue to develop as part of their work.

Description	Possible indicators	Component
High-performance computing clusters High-performance computing (HPC) clusters are the backbone of the compute infrastructure required for AI workloads. Measuring this infrastructure and its properties could provide insights into available national AI compute supply and demand.	**Hardware** specifications (quantitative and qualitative, like the specifications of a Top500 entry), and data centre infrastructure, such as: • Memory • Processors, co-processors, and cores • Power consumption • Number of national data centres and their ownership (i.e., public vs. private sector; domestic vs. international owned) **Performance** on HPC and ML benchmarks (i.e., based on AI application scenarios) such as: • Linpack benchmark score from the Top500 (a measure of a system's floating-point computing power) • Graph500 score (a rating of supercomputer systems focused on data-intensive loads) • MLCommons score (for machine learning workloads, based on application scenarios) **Utilisation:** • Utilisation rate of high-performance computing clusters for AI (i.e., available supply vs. average used supply) • Number of high-performance computing users and their affiliation (i.e., public vs. private sector) **Cost:** • Total cost of HPC clusters (i.e., capital expenditures vs. operations; private vs. public ownership; domestic vs. international)	Availability (supply) Use (demand) Access (cost) Security & Sovereignty (location, ownership)
Compute demand of AI systems Measuring the demand of AI systems is important to learn about the needs of the required compute infrastructure. Learning	• Estimate of average compute demand according to key stages of the AI system lifecycle, such as compute used for training and inferencing (i.e., per application such as natural language processing, computer vision etc.) • Estimate of key trends such as average data sets size and number of model parameters over time • Estimate of trends in the performance of AI hardware (i.e., peak performance in FLOPS, memory capacity, energy usage)	Use (demand)

about trends and developments can inform future considerations.	• Qualitative analysis of local and cross-border regulations in data sharing and privacy • Expected throughput (i.e., number of inferences) for AI applications per application domain (i.e., natural language processing, computer vision etc.) • Affiliation of AI system creator of state-of-the-art AI systems (i.e., public or private sector)	
Talent and skills Talent and skills enable compute infrastructure to be used effectively. Learning about the talent and skills landscape can inform compute policy decisions and direct investments.	• Prevalence of required skills in employment databases (e.g., LinkedIn) including key words and formal certifications • Number of students enrolled in relevant degrees (e.g., undergraduates, graduate students, and doctoral programs, also disaggregated by gender if possible) • Size of the AI compute research (industry and academia) community and amount of available funding in sectors to investigate skill and worker transfer • Number of yearly expected graduations in relevant degrees (also disaggregated by gender if possible) • Number of private service providers offering training of relevant skills	People (skills, training, diversity)
Private cloud providers Private cloud providers are key for researchers and industry to access AI compute. Learning about their setup, costs, etc. could inform the requirements for public cloud and HPC infrastructure.	• Percentage of market share of cloud providers for AI compute compared to, e.g., the capacity provided by academic HPC initiatives • Costs for renting AI compute via the cloud over time i.e., cost per hour or cost per FLOPs	Availability (supply) Access (cost)
National AI plans National AI plans provide insights into AI compute demand and needs based on a country's context and national objectives.	• Number of mentions of AI compute initiatives in national AI plans, such as in policies or national AI strategies • Qualitative analysis of AI compute initiatives in national AI plans, such as in policies or national AI strategies	Policy (law, regulation, strategy)
Public-sector spending on AI compute Public spending on AI compute can provide insights into current and future investments.	• Percentage of public sector spending on compute infrastructure (per relevant ministry) • Percentage of public sector spending on "on premise" vs. cloud compute • Number of international HPC programs with domestic participation • Number and quality of partnerships with global and regional commercial cloud providers, such as private-public partnerships	Availability (supply) Use (demand) Policy (law, regulation, strategy)
Private-sector spending on AI compute With AI increasingly used and developed in the private sector, measuring private-sector spending on AI compute can inform the state of current and future supply.	• Percentage spending of the private sector (per business sector) on compute infrastructure • Differences in compute spending by enterprise size (i.e., large business vs. small-and-medium enterprises)	Availability (supply) Use (demand)
Supply chains The resilience of AI compute supply chains is crucial to maintain sufficient AI compute capacity over time and in the face of possible economic and environmental shocks.	• Qualitative insights into and risks of the supply of the most prominent providers of compute hardware, especially for AI compute • Qualitative insights into risks and fragility (such as natural disasters and geopolitics) of supply chains for critical AI compute components	Security & Sovereignty (supply chains)

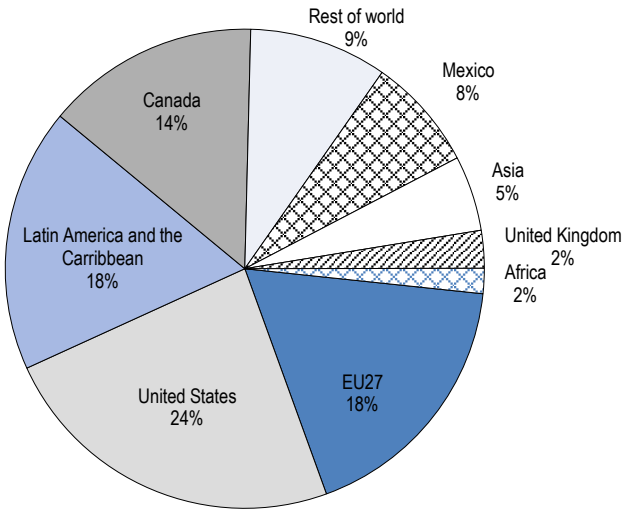
Annex D. Survey results on AI compute

Figure D.1. Survey respondents by sector



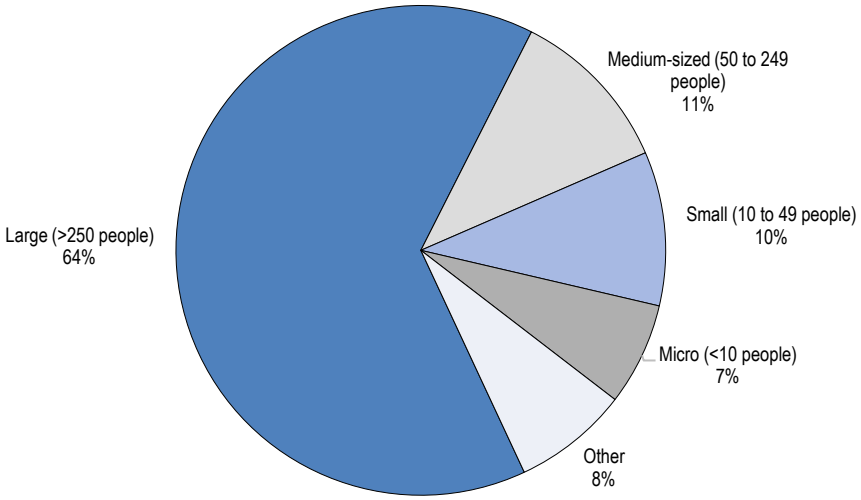
Note: Of the 118 respondents who partially or fully completed the survey, 117 respondents answered this question.
Source: OECD.AI Expert Group on AI Compute and Climate survey on measuring AI compute (March-April 2022)

Figure D.2. Geographic distribution of survey respondents



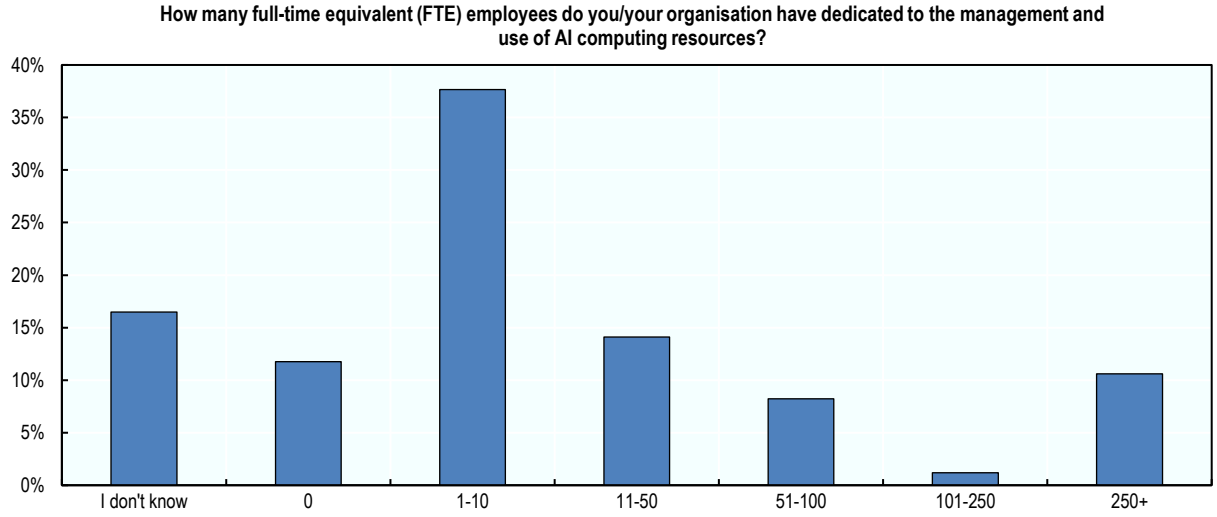
Note: Of the 118 respondents who partially or fully completed the survey, 118 respondents answered this question.
Source: OECD.AI Expert Group on AI Compute and Climate, survey on AI compute (March-April 2022)

Figure D.3. Organisation or enterprise size of survey respondents



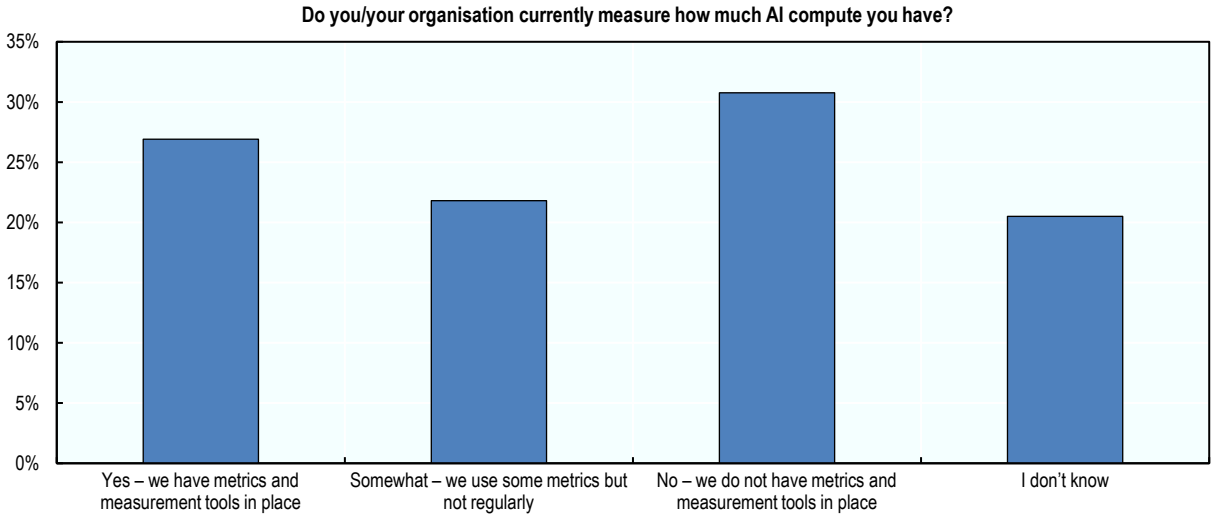
Note: Of the 118 respondents who partially or fully completed the survey, 118 respondents answered this question. According to the OECD (2022[50]), small and medium-sized enterprises (SMEs) employ fewer than 250 people. SMEs are further subdivided into micro enterprises (fewer than 10 employees), small enterprises (10 to 49 employees), and medium-sized enterprises (50 to 249 employees). Large enterprises employ 250 or more people.
Source: OECD.AI Expert Group on AI Compute and Climate survey on measuring AI compute (March-April 2022)

Figure D.4. Full-time equivalent (FTE) employees dedicated to the management and use of AI computing resources



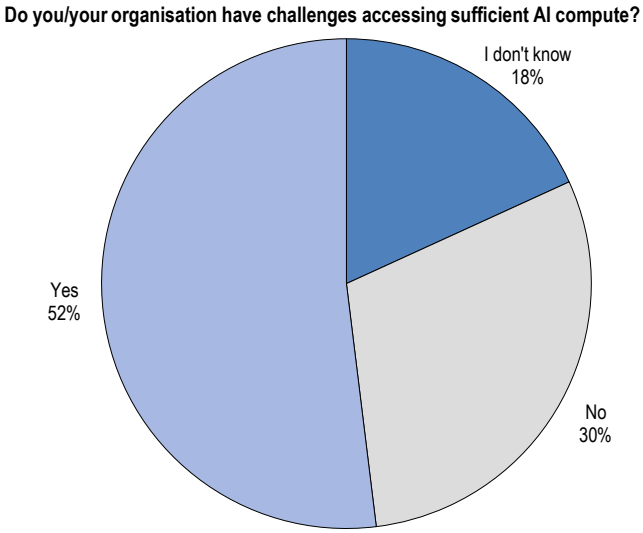
Note: Of the 118 respondents who partially or fully completed the survey, 85 respondents answered this question.
Source: OECD.AI Expert Group on AI Compute and Climate survey on measuring AI compute, March-April 2022

Figure D.5. Measurement of AI compute



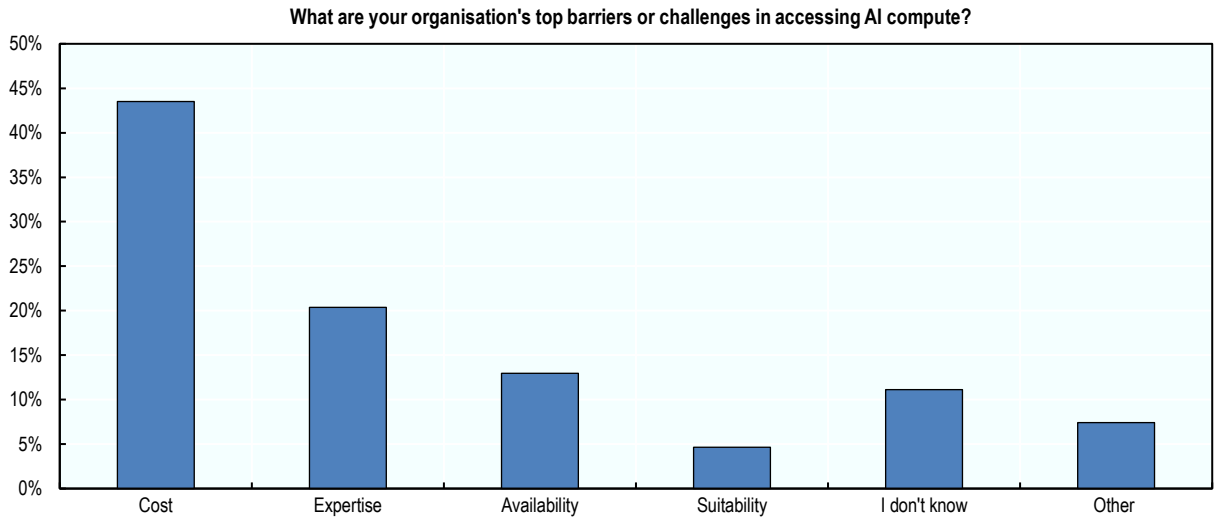
Note: Of the 118 respondents who partially or fully completed the survey, 78 respondents answered this question.
Source: OECD.AI Expert Group on AI Compute and Climate survey on measuring AI compute, March-April 2022

Figure D.6. Challenges accessing sufficient AI compute



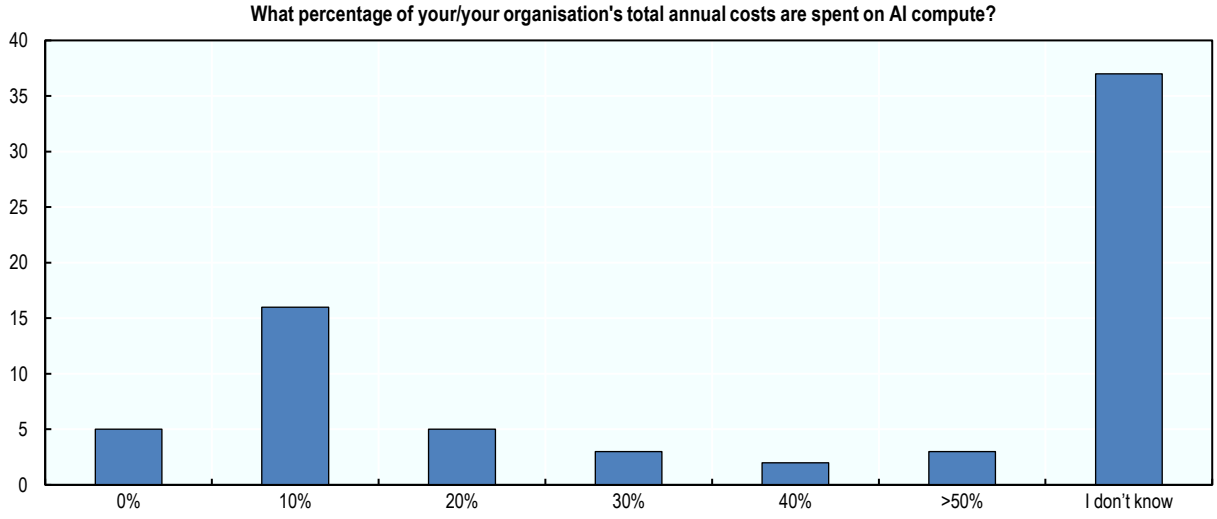
Note: Of the 118 respondents who partially or fully completed the survey, 77 respondents answered this question.
Source: OECD.AI Expert Group on AI Compute and Climate survey on measuring AI compute, March-April 2022

Figure D.7. Top barriers or challenges to accessing AI compute



Note: Of the 118 respondents who partially or fully completed the survey, 108 respondents answered this question.
Source: OECD.AI Expert Group on AI Compute and Climate survey on measuring AI compute, March-April 2022

Figure D.8. Cost allocation to AI compute



Note: Of the 118 respondents who partially or fully completed the survey, 77 respondents answered this question.
Source: OECD.AI Expert Group on AI Compute and Climate survey on measuring AI compute, March-April 2022

Annex E. Expert group co-chairs, members and observers, February 2023

Name	Title	Organisation	Group / Delegation
Ahuactzin, Juan Manuel	Research & Development Director	ProMagnus Company	Business
Aranda, Luis	Policy Analyst	OECD	Secretariat
Aristodemou, Leonidas	Analyst	OECD	Secretariat
Balasiano, Aviv	VP and Head of the Division	Technology Infrastructure in the Israeli Innovation Authority	Israel
Barrett, Gregg	CEO	Cirrus AI	Business
Bertrand, Arnaud	Chief Technical Officer and Senior Fellow	ATOS	Business
Bouvry, Pascal	Co-CEO	LuxProvide	Business
Caira, Celine	Economist/Policy Analyst	OECD	Secretariat
Cardoso Emediato de Azabuja, Eliana	General-Coordinator of Digital Transformation	Ministry of Science, Technology and Innovation	Brazil
Davidson, Landon	AI/ML Business Development	NVIDIA	Business
Clark, Jack	[ONE AI Chair] Co-founder	Anthropic	Business
Elison, David	Senior AI Data Scientist	Lenovo	Business
Escobar Silva, Maria Jose	Associate Professor	Universidad Técnica Federico Santa María	Civil Society and Academia
Escobar, Rebeca	Head of Studies Center	Federal Telecommunications Institute	Mexico
Fernández Gómez, Liliana	Advisor	Digital Development Directorate - National Planning Department	Colombia
Formica-Schiller, Nicole	Board Member	German AI Association (KI-Bundesverband)	Germany
Frankle, Jonathan	Chief Scientist / Assistant Professor of Computer Science	Mosaic ML and Harvard University	Civil Society and Academia
Garg, Arti	Chief Strategist, AI Solutions	Hewlett Packard Enterprise	Business
Gibson, Garth	Chief Executive Officer (former)	Vector Institute for AI	Civil Society and Academia
González Fanfalone, Alexia	Economist, Communication Infrastructures and Services Policy Unit	OECD	Secretariat
Heath, Tamsin	[ONE AI Chair] Deputy Director, Economic Security	Department for Digital, Culture, Media and Sport (DCMS)	United Kingdom
Heim, Lennart	Researcher	Centre for the Governance of AI	Civil Society and Academia
Hodes, Cyrus	Co-Founder	World Climate Tech Summit	Civil Society and Academia
Holoyad, Taras	Standards expert	Federal Network Agency for Electricity, Gas, Telecommunications, Post and Railway	Germany
Hui, Chen	Assistant Chief Executive	Infocomm and Media Development Authority (IMDA)	Singapore
Janapa Redi, Vijay	Associate Professor	Harvard University John A. Paulson School of Engineering and Applied Sciences	Civil Society and Academia
Javoršek, Jan Jona	Head of Networking Infrastructure Centre	Jožef Stefan Institute	Civil Society and Academia
Kanter, David	Executive Director	MLCommons	Business
Kent, Suzette	Business Executive, Former Federal Chief Information Officer of the United States	Kent Advisory Services	Business

Khareghani, Sana	Head (former)	UK Office for AI	United Kingdom
Kirnberger, Johannes Leon	AI and climate expert	Consultant on AI and Climate - OECD	Consultant
Krüppel, Roland	Electronics and Autonomous Driving; Supercomputing	Federal Ministry for Education and Research	Germany
Lee, Jiwon	Policy Officer	Ministry of Technological Innovation and Digital Transition	Italy
Lohn, Drew	Senior Fellow	Georgetown University Center for Security and Emerging Technology	Civil Society and Academia
Luccioni, Sasha	Research Scientist	Hugging Face	Civil Society and Academia
Macoustra, Angus	CTO, Head of Scientific Computing	Commonwealth Scientific and Industrial Research Organisation (CSIRO)	Australia
Mangla, Utpal	VP and Senior Partner	IBM Global Business Services	Business
Matsuoka, Satoshi	Director	RIKEN Center for Computational Science	Japan
Moetzel, Ulrike	Economist/political scientist	Federal Ministry for Digital and Transport	Germany
Moretti, Lorenzo	Innovation Policy Coordinator to the Minister	Ministry of Technological Innovation and Digital Transition	Italy
Mujica, María Paula	Advisor on Digital Transformation, Management and Compliance	High Presidential Advisory Office	Colombia
Nolan, Alistair	Senior Policy Analyst	OECD	Secretariat
Ouimette, Marc-Etienne	Global Lead AI Policy	Amazon Web Services	Business
Parashar, Manish	Director, Office of Advanced Cyberinfrastructure (OAC)	National Science Foundation	United States
Parker, Lynne	Deputy CTO of the United States of America (former)	United States Administration	United States
Perset, Karine	Head of Unit, OECD.AI	OECD	Secretariat
Radwan, Sally	Chief Digital Officer	United Nations Environment Programme	Civil Society an Academia
Rao, Anand	Global AI Leader	PwC	Business
Roquet, Ghilaine	Vice President of Strategy and Planning	Digital Research Alliance of Canada	Civil Society and Academia
Sampaio Gontijo, José Gustavo	Director	Department of Digital Science, Technology and Innovation	Brazil
Stancavage, Jayne	Global Executive Director, Digital Infrastructure Policy	Intel	Business
Stogiannis, Dimitris	Head of the Research, Development and Innovation (RDI) Statistics Unit	National Documentation Centre	Greece
Strier, Keith	[ONE AI Chair] Vice President	NVIDIA Worldwide AI Initiative	Business
Tretikov, Lila	CVP, Deputy Chief Technology Officer	Microsoft	Business
Georgios Tritsaris	Researcher	Sectoral Scientific Council in Natural Sciences (NCRTI, Greece)	Civil Society and Academia
Tyldesley, Jennifer	Deputy Director, Economic Security (former)	Department for Digital, Culture, Media and Sport (DCMS)	United Kingdom
Vasilis, Bonis	Technical Manager, Team Leader, Senior Software Architect and Technical Coordinator of European Projects	National Documentation Centre	Greece
Velsberg, Ott	Chief Data Officer	Ministry of Economic Affairs	Estonia
Weber, Verena	Head of Communication Infrastructures and Services Policy Unit	OECD	Secretariat
Yeong, Zee Kin	Assistant Chief Executive	Infocomm Media Development Authority of Singapore	Singapore
Zagler, Martin	Danish Business Authority	Ministry of Industry, Business and Financial Affairs	Denmark

Note: Member biographies are available on OECD.AI