

MEASURING THE ENVIRONMENTAL IMPACTS OF ARTIFICIAL INTELLIGENCE COMPUTE AND APPLICATIONS THE AI FOOTPRINT

OECD DIGITAL ECONOMY PAPERS

November 2022 **No. 341**



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 28/09/2022.

This report was developed in consultation with experts involved in the Global Partnership on Artificial Intelligence's Working Group on Responsible AI. Its contents reflect the opinions of the GPAI Experts involved and do not necessarily represent the official views of GPAI Members.

Note to Delegations:

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

DSTI/CDEP/AIGO(2022)3/FINAL

This document, as well as any data and any map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area. The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

© OECD 2022

The use of this work, whether digital or print, is governed by the Terms and Conditions to be found at <http://www.oecd.org/termsandconditions>.

3

Review of existing and emerging data and measurement frameworks

This section takes stock of existing data and measurement frameworks related to the environmental impacts of AI compute and applications. It distinguishes between direct and indirect environmental impacts, as described in Figure 2. Analysis of direct environmental impacts is guided by the four stages of the AI compute resources lifecycle (Figure 3). The third stage of the lifecycle (AI compute operations) explores the impact of AI compute across the sustainability metrics of energy and water consumption, and GHG emissions. Finally, existing and emerging data and measurement frameworks related to positive and negative indirect environmental impacts resulting from the application of AI, are also explored outside this framework, along with a discussion on dual impacts.

3.1. Direct environmental impacts of AI compute

Most indicators and frameworks focus on the direct environmental impacts of AI compute, as opposed to indirect impacts resulting from AI applications. These direct environmental impacts occur along the AI compute resources lifecycle: (1) production, (2) transport, (3) operations and (4) end-of-life (Figure 3). Within this lifecycle, researchers and practitioners largely have focused on impacts from the operations stage where AI compute is used to train or deploy AI systems. While AI applications have both positive and negative impacts, the direct environmental effects of AI compute are largely negative in terms of GHG emissions and resource consumption, with compute infrastructure sometimes requiring large amounts of energy and other material inputs (Barteková and Börkey, 2022^[20]).

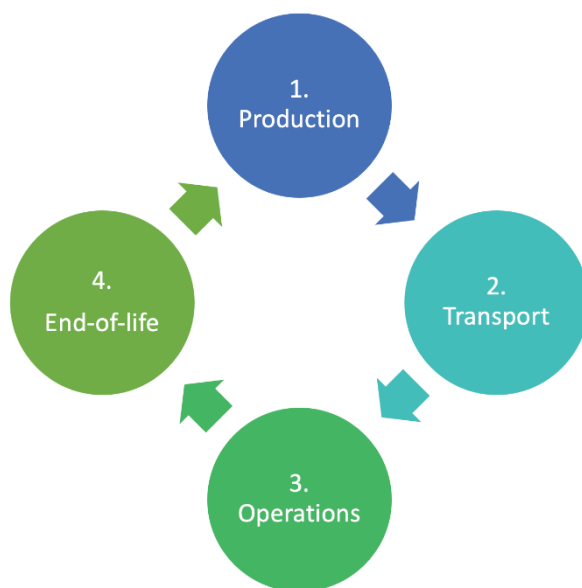
Guided by this framework, analysis of existing literature revealed that in the first, second and fourth stages of the AI compute resources lifecycle (production, transportation and end-of-life), most indicators and measurements relate to overall ICT equipment and hardware, which are not necessarily AI-specific. AI-specific indicators and measurements were only found for the third stage (operations), in particular related to energy consumption and GHG emissions (Rohde et al., 2021^[21]; Sustainability Index for Artificial Intelligence, 2022^[22]).

It is important to note that the AI resources lifecycle is a distinct framework from the AI system lifecycle as defined in the 2022 OECD Framework for the Classification of AI Systems, which defines the AI system lifecycle through the following phases: (1) planning and design; (2) collecting and process data; (3) building and using models; (4) verifying and validating the model; (5) deployment, and (6) operating and monitoring the system (OECD, 2022^[23]). These frameworks are distinct as they examine different aspects of AI, with the former focusing on environmental impacts AI compute from a natural resources perspective, and the latter focusing on the stages of AI development, application and monitoring.

In addition to the separate stages of the AI compute resources lifecycle, it is important to assess direct environmental impacts from their interconnection, to provide a comprehensive analysis (Wu et al., 2022^[24]). An example of such assessment is the Green Cloud Computing methodology from the German Environmental Agency, which looks at the following four impact categories: (1) abiotic resource depletion (i.e. use of minerals and fossil fuels); (2) cumulative energy demand (i.e. use of renewable and non-

renewable energy); (3) global warming potential (i.e. impact on climate change); and (4) water consumption (UBA, 2021^[25]). Full general lifecycle assessments (as defined in ISO standards 14040:2006 and 14044:2006) could be developed for individual AI compute components, such as the different products that make up AI compute stacks, to assess their environmental impact across lifecycle stages. However, applying existing tools like product lifecycle assessments can be challenging due to the distinct compute needs of AI systems requiring stacks of hardware and software that can vary significantly in their composition and include many different individual and highly specialised products.

Figure 3. Four stages of the AI compute resources lifecycle



Note: This framework is distinct from the AI system lifecycle defined in the 2022 OECD Framework for the Classification of AI Systems as: (1) planning and design; (2) collecting and processing data; (3) building and using models; (4) verifying and validating the model; (5) deployment and (6) operating and monitoring the system (OECD, 2022^[23]).

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

3.1.1. Production

AI compute production relies on the physical extraction and consumption of natural resources to build computing hardware, including computer chips, semiconductors, graphics processing units (GPUs) and central processing units (CPUs). There are many steps in the production of AI compute hardware and infrastructure, from mining, smelting and refining, to component manufacturing, such as semiconductor fabrication, and assembly. Environmental impacts along this value chain include soil contamination, deforestation, erosion, biodiversity degradation, toxic waste disposal, groundwater pollution, water use, radioactive waste and air pollution (Crawford, 2018^[26]). However, few frameworks and indicators differentiate between the computing resources used uniquely for AI and those used for other scientific, academic or industrial applications. Often, data is only available for total ICT infrastructure production and is not disaggregated for AI uses and applications.

Scientific literature reveals different estimates of how much the production stage adds to the environmental impacts of AI compute. Often, production impacts are overlooked due to difficulties in attribution (Henderson et al., 2020^[27]), with some scholars arguing that lifecycle assessments would be “impractical at scale and would greatly reduce who can use the method” (Lannelongue et al., 2021^[28]). Another reason given for excluding the production stage of AI compute from analysis is its presumed insignificance

compared to operational impacts (i.e. stage three of the AI compute resources lifecycle) (Siddik, Shehabi and Marston, 2021^[29]). Others underscore that adopting renewable energy procurement and improving energy efficiency in data centres could result in production representing a higher portion of AI compute's environmental footprint over time (Gupta et al., 2020^[30]). Non-operational impacts, such as those resulting in the production stage, will thus become relatively more important (IEA, 2021^[16]). At present however, many AI compute providers only report operational environmental impacts, excluding environmental impacts from production, which are difficult to estimate.

One estimate of the carbon footprint from the production of data centres globally is 20 megatons of carbon-dioxide-equivalent (CO₂e), representing about 15% of total data centre GHG emissions in 2015 (Malmodin and Lundén, 2018^[31]). The company Meta estimates that GHG emissions from the production of their data centres are around 30% of total company emissions for 2022 (Wu et al., 2022^[24]). But as data centres shift to carbon-free energy sources, their long-term operation will account for fewer emissions and some estimates project the share resulting from data centre production could eventually increase to over 80% (Gupta et al., 2020^[30]).

Value-chain GHG emissions arising up- and down-stream of an entity's immediate activities (known as "Scope 3 emissions"²) can account for a large proportion of the carbon footprint of cloud computing vendors. According to Meta, their share of value-chain emissions compared to total GHG emissions rose from 44% in 2017 to 99% in 2020 (Meta, 2021^[32]), resulting in a "significant embodied carbon cost paid upfront" for new AI-driven system components for data centres (Wu et al., 2022^[24]). Similarly, the share of value-chain GHG emissions according to Google increased from 45% in 2016 to over 90% in 2020 (Google, 2022^[33]). These increases in value-chain emissions are likely due to higher clean-energy procurement, which lowers the proportion of emissions from energy use. It should be noted that a carbon footprint analysis alone does not reflect the full environmental impact of AI compute production, including the mining of raw materials that can be environmentally damaging and carry significant supply and human rights risks (Crawford, 2021^[34]). The environmental and social impacts of extracting raw materials like cobalt, palladium, copper, lithium and aluminium are rising on the agenda of policy makers (European Parliament, 2021^[6]) and industry networks (Sustainable Infrastructure Alliance, 2022^[35]).

Indicators

Metrics to consider at the production stage of the AI compute resources lifecycle:

- GHG emissions from production, in metric tons of CO₂e
- the carbon intensity of production methods, in metric tons of CO₂e per unit (e.g. per person or dollar of revenue)
- the share of renewable energy used in production
- the share of recycled or renewable materials used in production

3.1.2. Transportation

As with production, the environmental impacts of transporting AI compute hardware are difficult to disaggregate from those of transporting other ICT hardware, and even from the transportation of non-ICT goods in general. Transport activities can generate various negative environmental impacts, such as air pollution, oil spills, toxic-waste discharges and acoustic pollution (Crawford, 2018^[26]; OECD, n.d.^[36]). Global distribution, freight transportation, handling and storage depend on fossil fuels, with road-freight vehicles responsible for 2.4 gigatons of CO₂e per year (around 6% of global energy-related emissions), and with shipping and aviation each responsible for about 1 gigaton of CO₂e per year (around 2.5% of global energy-related emissions) (IEA, 2021^[37]).

But only a small fraction of global transportation figures can be attributed to ICT hardware, and an even smaller portion can be attributed to the transportation of AI compute hardware, such as specialised GPU racks. The few estimates available put transport of AI compute hardware at less than 5% of total GHG emissions over an AI system's lifetime, with some estimates as low as less than 1% (Gupta et al., 2020^[30]). The differences in estimates could be attributed to the location of hardware production and the corresponding transport distance.

The decarbonisation and sustainability of transportation is a global undertaking. While transportation of AI compute hardware has a small impact compared to other lifecycle stages, opting for sustainable and environmentally responsible transportation methods should be further explored and considered.

Indicators

Metrics to consider at the transportation stage of the AI compute resources lifecycle:

- GHG emissions from transportation of AI compute hardware in metric tons of CO₂e
- the carbon intensity of the transport methods in metric tons of CO₂e per unit (e.g. per person or dollar of revenue)
- the share of low-carbon energy and/or renewable energy used in these transport methods

3.1.3. Operations

The environmental impacts of operating AI compute primarily relate to energy consumption, GHG emissions and water consumption, which can occur when an AI system is developed, such as through training, and deployed, for example by applying AI models to make predictions, recommendations or decisions (also known as inference). These three environmental impacts are further explored below.

Energy consumption

Compared to the production and transport stages of the AI compute resources lifecycle, energy consumption is well documented for global data centres and ICT at large. Although a range of estimates exist (Banet et al., 2021^[38]), researchers and institutions arrive at fairly consistent results. The IEA estimates global data centre electricity demand at 194 Terawatt Hours (TWh) in 2014, or 1% of global electricity demand (IEA, 2017^[39]). By 2020, due to large efficiency gains, that estimate had only risen to 200-250 TWh and remained at 1% of global electricity demand (IEA, 2021^[16]). Other experts estimate worldwide energy demand by data centres in a similar range: at 190 TWh in 2020, with most coming from cloud and hyperscale operators (Sönnichsen, 2021^[40]). Masanet et al. estimate demand at 205 TWh in 2018, or 1% of global electricity consumption, and just a 6% increase from 2010 despite a 550% increase in data centre compute demand over the same time period (Masanet et al., 2020^[41]).

Notably, efficiency improvements and the shift to large hyperscale data centres have offset much of the exponential growth in data centre services in the past decade, keeping the estimated energy use of data centres almost flat (IEA, 2021^[16]). Further efficiency could be achieved through specialised AI hardware and network architecture, such as breakthroughs in energy efficiency using ultra-low-energy artificial neurons (IEEE, 2022^[42]). However, with technological changes in ICT and the acceleration of digital transformation during the COVID-19 pandemic, this trend might vary in the future. Some estimate that data centres could account for 783 TWh by 2030, or around 2.6% of global electricity use which is estimated to reach 30 000 TWh in 2030 (Andrae, 2020^[43]).

At the national level, few governments track energy consumption and projected demand by ICT infrastructure and related AI compute resources within their borders. In Denmark, annual demand by data centres is expected to grow from near-0 in 2020 to 5 TWh by 2025 and 7.5 TWh by 2030 (Danish Energy

Agency, 2020^[44]). In Ireland, a European hub for data centre operators, electricity consumption by data centres increased 144% from 2015 to 2020, accounting for 11% of metred electricity consumed in the country in 2021 (Central Statistics Office, 2022^[45]). Median-demand scenarios estimate that this figure could rise as much as 23% by 2030 (EirGrid, 2021^[46]). In the Netherlands, data centre energy use grew from 1.6 TWh in 2017 to 2.7 TWh in 2019, representing 2.7% of national electricity supply (CBS, 2021^[47]). In the United States, data centres accounted for 70 TWh or 1.8% of total electricity consumption in 2014 (Shehabi et al., 2016^[48]). Estimates for total electricity consumption by data centres in the European Union (EU27) in 2018 range between 42 TWh (European Commission, 2022^[49]) and 54 TWh for EU27 countries in addition to the United Kingdom (Montevecchi et al., 2020^[50]), or 2-3% of overall electricity consumption.

A Greenpeace report puts Chinese data centre electricity consumption at 161 TWh in 2018, of which 75% came from coal-fired power stations (Greenpeace China, 2021^[51]). The Chinese State Grid Energy Research Institute reported that the People's Republic of China's (hereafter "China") data centre energy consumption at 200 TWh in 2020. However, the details of whether such estimates include 5G networks and bitcoin mining are unclear (Chinese State Council, 2021^[52]). If accurate, this could indicate that current estimates of global data centre energy consumption are lower-bound estimates, and real energy consumption could be higher (Hintemann, 2018^[53]).

Around the world, recent reports have pointed to the demand for data centres contributing to increased pressure on local energy grids. There are reports of new housing developments being halted due to the high electricity demands of data centres (Financial Times, 2022^[54]), while some jurisdictions consider a moratorium on new data centre construction due to strained national power supplies and energy grid constraints (The Times, 2022^[55]). Climate change and more frequent heatwaves are also reported to add stress on power grids and data centres, which can lead to outages (Google Cloud, 2022^[56]).

Many data centre operators report energy consumption, including total electricity consumption, energy efficiency and power usage effectiveness (PUE), renewable energy procurement, facility energy use, and renewable energy consumption (Apple, 2021^[57]; Meta, 2021^[32]; Google, 2022^[33]; Microsoft, 2021^[58]; Digital Realty, 2019^[59]; Equinix, 2021^[60]). ICT companies and large operators are major purchasers of renewable energy and account for almost half of global corporate renewable energy procurement (IEA, 2021^[16]). Worldwide associations like the Energy Efficient High Performance Computing Working Group and the Center of Expertise for Energy Efficiency in Data Centers, aspire to drive energy efficiency and share best practices (EEHPCWG, 2022^[61]; Center of Expertise for Energy Efficiency in Data Centers, 2022^[62]).

Several approaches and frameworks estimate energy consumption by machine learning models. These allow for a more granular assessment of dedicated AI models than do global estimates. The use of standard metrics for reporting efficiency, such as training time and hyperparameter sensitivity, could improve comparability between models (Strubell, Ganesh and McCallum, 2020^[63]). The Green500 list ranks the top 500 most powerful supercomputers from around the world according to their energy efficiency (Green500, 2021^[64]). Others propose Bayesian deep learning (Welling, 2018^[65]) and sustainable federated learning (Guler and Yener, 2021^[66]) to make AI more power-efficient. Future machine learning models will need to balance performance and efficiency as demand for AI compute grows with the creation of more complex and large AI models (Desislavov, Martínez-Plumed and Hernández-Orallo, 2021^[67]).

Box 3. Optimising energy use to reduce direct environmental impacts at DeepMind AI

Over 40% of DeepMind AI's data centre energy use comes from cooling systems. AI can help operators make data centres more efficient, decreasing energy consumption and resulting GHG emissions. DeepMind's reinforcement learning algorithm was applied at one of Google's data centres and is reported to reduce energy use by 30-40% compared to cooling systems without optimisation.

Source: (DeepMind, 2016^[68]; DeepMind, 2018^[69]; Zhang et al., 2017^[70])

Indicators

Energy consumption metrics to consider at the operations stage of the AI compute resources lifecycle:

- electricity consumption in TWh
- renewable electricity consumption in TWh
- Power Usage Effectiveness (PUE) for total facility power compared to ICT equipment power

Greenhouse gas emissions

The operational carbon footprint of AI compute relates directly to its energy consumption, often from non-renewable energy sources. The global ICT industry (including hardware such as televisions) is estimated to be responsible for 1.8-2.8% of global GHG emissions, while other calculations estimate it as high as 2.1-3.9% (Freitag et al., 2021^[71]). A more suitable estimate (excluding televisions) puts the ICT-industry figure at 700 metric tons of CO₂ in 2020, or 1.4% of global emissions (ITU, 2020^[72]).

Examining trends for ICT emissions, some researchers claim GHG emissions have plateaued as “footprints are significantly smaller than previously forecasted” (Malmudin and Lundén, 2018^[31]). Differences in estimates can be explained by underlying assumptions about the future global energy supply mix, as projections depend largely on the carbon intensity of electricity supply and the availability of renewable energy sources. The ITU and the World Benchmarking Alliance (WBA) estimate that the 150 largest digital companies account for 1.6% of global electricity use, of which 32% comes from renewable sources (International Telecommunication Union and World Benchmarking Alliance, 2022^[73]). In the United States, around 0.5% of national emissions are estimated to be from data centres (Siddik, Shehabi and Marston, 2021^[29]). More broadly, cloud and hyperscale data centres are estimated to account for 0.1-0.2% of global emissions (Kaack et al., 2022^[74]).

AI compute is estimated to account for only a fraction of these numbers. However, few estimates of AI's energy demand exist, with these estimates rarely differentiating between AI training and use (or “inference”) workloads. According to one study by Google, its overall energy use for machine learning workloads consistently represented less than 15% of total energy use over the three-year period from 2019 through 2021 (Patterson, 2022^[75]). Other estimates use customer spending to approximate the percentage of compute used between AI training and inference workloads. For example, a large cloud compute provider³ estimates that its enterprise customers spend between 7-10% of their total compute infrastructure expenditure on supporting AI and machine learning applications, broken down between training and use (“inference”) at approximately 3-4.5% for machine learning training and 4-4.5% for machine learning 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 energy use and associated GHG emissions, while shedding light on how impacts differ according to whether compute is used for AI training or inference.

Many of the world's largest data centre operators announced plans to power them with 100% renewable energy (Dhar, 2020^[76]). Amazon announced its intention to power all company operations with renewable energy by 2025 and become carbon-neutral by 2040 (Amazon, 2022^[77]). In 2021, it reported to source 85% of its energy from renewable sources across its business, including AWS (Amazon, 2022^[77]). Google reports on its emission inventory, compensation for emissions, and carbon intensity, matching 100% of annual electricity use with renewable energy since 2017. Annual electricity matching means purchasing renewable energy to equal annual energy consumption through power-purchase agreements. As annual electricity matching can still contain a significant share of non-renewable resources, Google committed to run on 24/7 carbon-free energy by 2030, meaning that every kilowatt-hour of electricity use is met with carbon-free electricity sources (Google, 2022^[33]).

Google also introduced a policy roadmap to achieve completely carbon-free energy (Google, 2022^[78]) and a system for “carbon-intelligent” compute management, which generates “carbon-aware” capacity estimates for its data centres (Radovanovic et al., 2021^[79]). Similarly, Meta reports on GHG emissions and carbon intensity and announced it has been running carbon-neutral operations as of 2020. Meta plans to achieve net-zero emissions throughout its supply chain by 2030 (Meta, 2021^[32]). It also implemented designs for “carbon-aware” data centres under its Carbon Explorer Framework (Acun et al., 2022^[80]). Microsoft reports on its emission inventory and carbon-neutrality offsets and aims to become carbon negative by 2030 (Microsoft, 2021^[58]).

Major cloud computing vendors offer tools for their customers to estimate carbon-footprints based on their cloud compute usage. AWS introduced a carbon-footprint tool to help customers understand and forecast emissions of cloud services (Amazon Web Services, 2022^[81]). This includes a shared-responsibility model in which they provide sustainable infrastructure, efficient cloud usage and workloads while customers are encouraged to run their models efficiently (Amazon Web Services, 2021^[82]). Google provides carbon-footprint reports based on the Greenhouse Gas Protocol carbon accounting standards (Google Cloud, 2022^[83]). Microsoft offers an emissions impact dashboard showing direct and indirect GHG emissions of its cloud services (Microsoft, 2022^[84]). Third-party providers also offer free and open-source tools to estimate cloud GHG emissions (Cloud Carbon Footprint, 2022^[85]). It is important to note that these customer-facing tools only estimate operational emissions from AI training and use, and not embodied emissions from the entire AI compute resources lifecycle, such as emissions created during hardware and infrastructure production, transportation and end-of-life.

A number of researchers put forward tools and frameworks to estimate the carbon impact of AI models, such as the Machine Learning Emissions Calculator (Lacoste et al., 2019^[86]), a free online tool called Green Algorithms (Lannelongue et al., 2021^[28]) and a software package that can be integrated into a Python codebase (CodeCarbon, 2022^[87]). One framework, called the experiment-impact-tracker, provides an interface for real-time tracking of GHG emissions and explores how regional energy-grid differences can reduce emissions by up to 30% using low-carbon energy sources (Henderson et al., 2020^[27]). Researchers also put forward best practices, including the use of pre-trained models and transfer-learning, where appropriate; efficient machine learning model architectures; processors optimised for machine learning, cloud computing and usage optimisation; and shifting operations to data centres with low-carbon energy availability (Patterson et al., 2021^[88]).

Indicators

GHG emissions metrics to consider at the operations stage of the AI compute resources lifecycle:

- GHG emissions in metric tons of CO₂e
- carbon intensity in metric tons of CO₂e per unit (e.g. per person or dollar of revenue)
- carbon usage effectiveness (CUE): the ratio of total CO₂e emissions caused by total data centre energy consumption to the energy consumption of ICT equipment

Water consumption

While many discussions about sustainable compute revolve around energy efficiency and zero-carbon operations, freshwater consumption is a resource with major, often-overlooked environmental impact (Heslin, 2016^[89]). AI compute hardware and infrastructure consume water in two major ways: (1) directly, for cooling and (2) indirectly through water use for electricity generation. The production stage of AI compute, such as semiconductor fabrication, can also use large amounts of water. Compared to operational energy use and GHG emissions, water consumption is poorly understood. Only about 33-50% of data centre operators compile and report water-use metrics (Mytton, 2021^[90]; Uptime Institute, 2021^[91]; Google Cloud, 2022^[83]) such as water withdrawal minus water consumption, or water returned to the local water system following use (Microsoft, 2021^[58]; Meta, 2021^[32]; Siddik, Shehabi and Marston, 2021^[29]).

In the United States, data centres are estimated to account for less than 1% of total water consumption (Mytton, 2021^[90]). However, they compete with users such as hospitals or agricultural production. The United States data centre industry draws directly and indirectly from 90% of national watersheds and is one of the top ten water-consuming industries in the country. Data centres often cluster in similar geographic areas and many rely on scarce water supplies, particularly in the western United States (Siddik, Shehabi and Marston, 2021^[29]).

Water consumption increasingly features in debates around the sustainability of AI compute. European industry associations for data centres and cloud infrastructure list water conservation as a priority (Climate Neutral Data Centre Pact, 2022^[92]). Data centres based on liquid cooling could also recover the excess heat for on-site GHGs, nearby buildings or local heating, in what some call an Organic Data Centre (ODC) approach (Karnama, Haghighi and Vinuesa, 2019^[93]).

Box 4. Efficient water and energy use to minimise environmental impacts: Lefdal Mine Datacenter

Located inside a former mine in the Sogn og Fjordane region of Norway, the Lefdal Mine Datacenter offers a case study in sustainable data centre architecture that minimises the direct environmental impacts of compute. Using 100% renewable hydropower electricity, the data centre guarantees a PUE of 1.15, close to the generally accepted ideal PUE of 1.0 (ratio of total energy used to the energy delivered to computing equipment). Notably, Lefdal Mine uses very little water in its operations. It also reuses the heat generated by cooling solutions based on the supply of cool sea water. As a signatory of the Climate Neutral Data Centre Pact and the iMasons Climate Accord, its data centre is on track to become fully carbon-neutral.

While its unique location and geographic advantages are hard to replicate, it is an example of the Norwegian Data Centres Strategy goal to become “the world’s most sustainable data centre nation” by highlighting the potential of renewable energy, low water usage and ways to re-use excess heat.

Source: (Lefdal Mine Datacentre, n.d.^[94])

Indicators

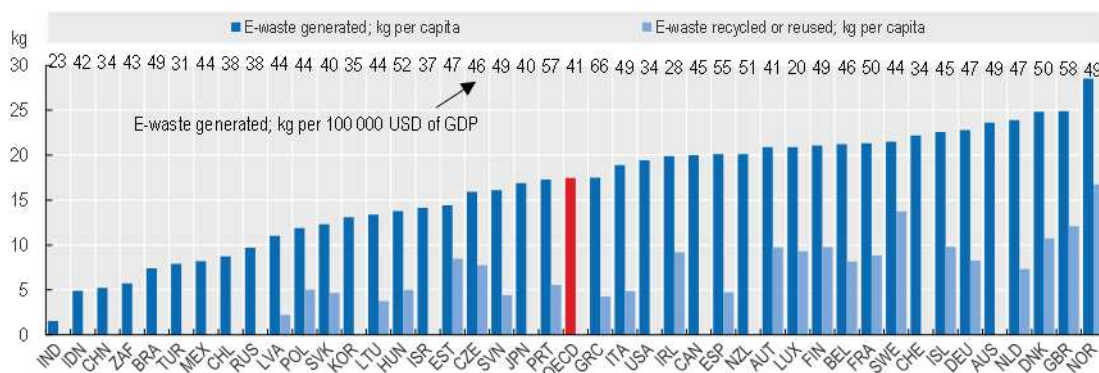
Water-consumption metrics to consider at the operations stage of the AI compute resources lifecycle:

- water withdrawal in cubic meters
- water consumption in cubic meters
- water discharge (quality) in cubic meters
- water usage effectiveness (WUE) in litres per kilowatt-hour
- water withdrawal intensity in cubic meters per unit (e.g. per person or dollar of revenue)

3.1.4. End-of-life

The resources lifecycle of AI compute ends with recycling or disposing of electronic waste (e-waste). The collection, shipping, recovering and disposal of AI compute hardware has environmental and social impacts such as air pollution, acidic and radioactive waste, groundwater pollution, and more (Crawford, 2018^[26]). Much of global e-waste disposal is conducted in developing countries, adding to their environmental and social challenges (Forum, 2019^[95]). The OECD tracks e-waste, defined as all items of electric or electronic equipment discarded by their owner, and national recycling levels (OECD, 2019^[96]). There is a significant gap between e-waste generated and recycled or reused (Figure 4).

Figure 4. E-waste generation and recycling or reuse, 2016



Note: This figure does not distinguish between AI and non-AI e-waste. The figure covers six waste categories: (1) temperature equipment; (2) screens, monitors; (3) lamps; (4) large equipment; (5) small equipment; and (5) small IT and telecommunication equipment. Disaggregating these categories could better estimate AI compute-related impacts.

Source: (OECD, 2019^[96]), see Statlink: <https://doi.org/10.1787/888933931086>

As in the production and transport stages of the AI compute resources lifecycle, most indicators and measurements of end-of-life impacts relate to overall ICT equipment and hardware. ICT infrastructure accounts for about 12 million tons, or 25% of total global electronic waste (Sustainable Infrastructure Alliance, 2022^[35]). Increasingly, used AI compute equipment is also sold, repurposed and fed into closed-loop supply chains that re-use or recycle materials (IT Renew, 2022^[97]). The European Commission launched the Sustainable Products Initiative, designed to make ICT equipment more durable, reusable, repairable and recyclable, and proposed a digital product passport to track hardware along its lifecycle. Further efforts could be made to incorporate principles of circular design, extended-life hardware, and recycling for AI related equipment (Sustainable Infrastructure Alliance, 2022^[35]).

Indicators

Metrics to consider at the end-of-life stage of the AI compute resources lifecycle:

- electronic waste in metric tons
- recycling rate
- electronics disposal efficiency (EDE) percentage
- percentage of electronic waste sent to landfills

3.1.5. Positive direct environmental impacts of AI compute

Some scholars propose the term “green AI” to describe AI systems that do not increase (and ideally, decrease) the environmental costs associated with compute. This contrasts with “red AI”, where accuracy is obtained through the use of massive AI compute resources with little regard for the environmental impact (Schwartz et al., 2019^[98]). Since even green AI consumes resources, it still has negative impacts on the environment. In only a handful of cases do positive direct environmental impacts result from AI compute, which are described below. These could be further analysed, expanded and emulated by AI compute providers.

Data centres produce large amounts of excess heat, typically considered “low-grade energy”. This energy usually cannot be repurposed for other activities as the temperatures are too low, typically below 35 degrees Celsius. Instead, excess heat is often directed into cooling towers. Several methods have been proposed to recover this heat, for example by combining the operation of a data centre and an onsite greenhouse or transferring it to local energy networks (Karnama, Haghighi and Vinuesa, 2019^[93]; ReUseHeat, 2017^[99]).

In terms of water consumption, several cases show the quality of wastewater released from data centres to actually be higher than when it was drawn from the source (Siddik, Shehabi and Marston, 2021^[29]). However, this wastewater often must be treated after use, which consumes electricity and produces emissions, rendering the net sustainability impacts unclear.

3.2. Indirect environmental impacts of AI applications

The indirect environmental impacts of AI compute, which result through the application of AI, are a growing area of research and analysis. The application of AI and its structural and behavioural effects can cause both positive and negative indirect environmental impacts. However, given the multiplicity of AI applications and their indirect impacts’ diffuse nature, these are much harder to quantify than direct impacts. To date, assessments of indirect environmental impacts are largely qualitative, but their potential large-scale effect on climate action and planetary health provide a compelling case for the further development of analytical and measurement frameworks.

3.2.1. Negative indirect environmental impacts of AI applications

AI applications can cause negative environmental impacts. For example, AI applications can exacerbate the negative environmental impacts of the mining, extractive and manufacturing sectors, where advanced AI applications can be used upstream for finding and extracting minerals or fossil fuels, midstream for transport and material storage, and downstream for product refining. While AI can increase efficiencies that support sustainability efforts, these applications can also work to increase net GHG emissions instead

of lowering them. For example, AI recommender systems used in e-commerce could increase consumption in unsustainable ways. Companies have also been reported to use the cloud to shift IT-related emissions from mandatory reporting requirements (e.g. on-premise) to voluntary reporting categories (e.g. outsourced cloud services) to “hide greenhouse gas emissions in the cloud” (Mytton, 2020^[100]).

AI applications can also have complex, systemic effects on the environment and human behaviour. For instance, researchers observe that efficiency gains can be offset by “rebound effects” that cancel out positive sustainability impacts (Paul et al., 2019^[101]). Rebound effects such as the “Jevons Paradox” occur when efficiency gains through technological progress are offset by acceleration in resource consumption (Giampietro and Mayumi, 2018^[102]). Some researchers also observe a low willingness among consumers to pay for energy-efficient AI applications (König, Wurster and Siewert, 2022^[103]). This led some researchers to point out the limitations of relying on efficiency increases alone, asserting that the compute demands of AI in its current form make future progress “economically, technically, and environmentally unsustainable” (Thompson et al., 2020^[14]). Applying behavioural insights (e.g. recognition of the effect of different consumer biases on behaviour) to this area may assist in designing more effective policies (e.g. labelling schemes or mandatory information disclosures) that nudge consumers towards more energy efficient AI applications (OECD, 2017^[104]).

3.2.2. Positive indirect environmental impacts of AI applications

Rolnick et al. highlight the areas in which AI applications can have a positive impact on climate action. These include electricity systems, transportation, buildings and cities, industry, farms and forests, GHG removal, climate prediction, societal impacts, solar geoengineering, individual and collective action, education and finance (Rolnick et al., 2022^[105]). Energy system operators increasingly leverage AI-supported digital twin simulations to optimise energy systems and other environmental parameters. A digital twin can be defined as “a digital representation of a real-world entity or system...that mirrors a unique physical object, process, organization, person or other abstraction” (Gartner, n.d.^[106]). On a larger scale, AI is leveraged in initiatives such as the European Space Agency’s Destination Earth project to digitally model the Earth’s environmental systems (Box 5).

Box 5. Digital twin for forecasting and resilience against climate change at Destination Earth

Understanding the interactions between human activities and natural phenomena is key to combating climate change and the biodiversity crisis. As part of the European Commission’s Green Deal, Destination Earth (DestinE) is developing a digital twin of the Earth, supported by AI capacity and the EU’s High-Performance Computing Joint Undertaking. Implemented by the European Space Agency, the European Centre for Medium-Range Weather Forecasts (ECMWF) and the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT), Destination Earth will enable dynamic Earth simulations, improve prediction capabilities and inform European environmental policy-making.

Source: (European Commission, 2022^[107]; European Commission, 2022^[49]; ECMWF, 2022^[108])

Several consulting firms put forward estimates of the environmental impacts of AI applications, although their methodologies vary and often are not publicly disclosed (Kaack et al., 2022^[74]). Boston Consulting Group (BCG) estimate that AI applications could help eliminate 2.6 to 5.3 gigatons of GHG emissions (5-10% of total emissions) and generate USD 1.3 to 2.6 trillion in value in revenues and cost savings through corporate sustainability by 2030 (BCG, 2021^[109]). Capgemini estimated in 2020 that AI could reduce GHG emissions by 16% and improve power efficiency by 15% by 2025 (Capgemini, 2020^[110]). According to PricewaterhouseCoopers (PwC), AI could increase global GDP 4.4% and reduce GHG emissions by 4%, or 2.4 gigatons, by 2030 (PWC, 2018^[111]).

AI can enable and contribute to the 17 UN Sustainable Development Goals (SDGs), including climate action (SDG 13), life below water (SDG 14) and life on land (SDG 15) (Vinuesa et al., 2020^[112]). Examples include facilitating climate analysis and forecasting, promoting energy conservation, improving GHG absorption and carbon storage, as well as decarbonising carbon-intensive sectors such as transport and agriculture (AI4SDGs, 2022^[113]). Beyond climate action, AI applications can also be used for biodiversity conservation, healthy oceans, water security, clean air and weather and disaster resilience (World Economic Forum, 2018^[114]). For example, AI applications have been leveraged to prevent wildlife destruction by supporting local conservation efforts (Gomes, 2019^[115]) and could help cities detect costly water leaks in municipal water systems (Cody, Harmouche and Narasimhan, 2018^[116]).

Use cases and tools have emerged in recent years to promote the use of AI for environmental sustainability. For instance, AI can be applied towards improved precipitation “nowcasting” — the high-resolution forecasting of precipitation for weather-related decision-making (Ravuri et al., 2021^[117]) — and long-term forecasts and sustainability insights for agricultural production, which can increase climate resilience and agricultural productivity (ClimateAI, 2022^[118]). AI is starting to be more widely used by organisations, municipalities and regions to measure, simulate and reduce the environmental footprint of supply chains (Barteková and Börkey, 2022^[20]). AI can also be leveraged to scale the transition to a resource-efficient and circular economy, “decoupling economic activity from natural resource use and its environmental impacts” (Barteková and Börkey, 2022^[20]). Finally, AI applications can accelerate scientific research and breakthroughs in the development and deployment of sustainable technologies. For example, researchers predict that a type of AI called deep reinforcement learning could accelerate the development of nuclear fusion, a potentially transformative carbon-free technology for the world’s energy demands (Degraeve et al., 2022^[119]).

It is important to note that most benchmarks, frameworks and impact assessments for the indirect environmental impacts of AI compute remain high-level and qualitative. Some researchers put forward approaches for measuring the environmental impacts of ICT overall, such as a methodology for assessing the effects induced by ICT services (Bergmark et al., 2020^[120]), the assessment of indirect environmental effects of digitalisation based on a time-use perspective (Bieser and Hilty, 2018^[121]), and an ICT-for-sustainability framework (Hilty and Aebischer, 2015^[122]) that was extended specifically to machine-learning-related GHG emissions (Kaack et al., 2022^[74]). Nevertheless, AI-specific methodologies for indirect environmental impacts remain rare – a key gap and limitation identified by this report.

Box 6. Tracking global greenhouse gas emissions through Climate TRACE

National GHG emission tracking historically relies on self-reporting and bottom-up estimates. The Climate TRACE coalition of AI analytics organisations leverages over 300 satellites and 11 000 sensors to quantify emission sources through AI algorithms. Emission sources include oil and gas production, shipping, aviation, forestry and more. Climate TRACE is relevant in more than 100 countries that lack accurate emission inventories of the past five years.

Source: (GPAl, 2021^[9]; Climate Trace, 2022^[123]; Carbon Tracker, 2020^[124])

Indicators

Key consideration for the measurement of negative and positive indirect environmental impacts associated with AI compute applications include:

- Impact assessment for net environmental cost or benefit of using an AI application (i.e. the energy/water/carbon/resources saved compared to energy/water/carbon/resources used)

3.3. Dual impacts

Like many technologies, the use of AI compute resources and applying AI itself can have both positive and negative effects, providing an opportunity to harness the benefits while minimising the costs to achieve a net positive environmental impact (Cowls et al., 2021^[125]). Machine learning and cloud computing can “help reduce stress on the environment in specific domains”, but “raise concerns about environmental and climate impact”, including use of energy and natural resources, and waste disposal (EPFL International Risk Governance Center, 2022^[126]). Addressing these dual impacts is an area for further study, requiring policy frameworks, tools and measurements to assess the net impacts of AI on the environment, developed through inter-disciplinary efforts by scientists, policy makers, technologists and others (Stein, 2020^[127]).