PUBLICATIONS · ACADEMIC REPORT

AI'S CARBON FOOTPRINT

THE ENVIRONMENTAL IMPACT OF ARTIFICIAL INTELLIGENCE

Ethical Tech CoLab
Professor Yorke Rhodes III
May 2025

Alexandra Du, Elizabeth Matthews, Emily Harrington, Hannah Zhao, Jennifer Hofmann, Natasha Nagarajan, Renata Gladkikh, and Smita Samanta

> AI's increasing environmental costs — energy-intensive training and inference, data centers, and hardware — require urgent action. A combination of sustainable AI practices, policy interventions, and technological innovations can significantly reduce AI's ecological footprint without hindering its development and role in advancing society.

KEY FIGURES

1,287 MWH

to train GPT-3 — the annual energy use of ~130 U.S. households

552 → 24 T

CO_2 to train GPT-3 on a coal grid vs. a renewable-powered grid

240–340 TWH

electricity the world's data centers already draw each year

10×

the emissions of a query in coal-heavy Wyoming vs. hydro Quebec

01

INTRODUCTION

Artificial intelligence (AI) is transforming industries and tackling global challenges, including climate change. However, AI's rapid expansion comes with a significant but often overlooked environmental cost. The development and deployment of AI models, particularly large-scale systems, consume vast amounts of energy (Harvard Business Review, 2024). This contributes to carbon emissions, water consumption, and electronic waste (Earth.org, 2025). As AI adoption grows, understanding and mitigating its ecological footprint is critical. This paper examines the environmental impact of AI, focusing on its energy demands, data center infrastructure, and hardware requirements. It also explores mitigation strategies, ethical considerations, and policy frameworks to promote sustainable AI development. While AI presents significant technological advancements for society, its increasing environmental costs seen through energy-intensive training and inference, data centers, and hardware, require urgent action. This paper argues that a combination of sustainable AI practices, policy interventions, and technological innovations can significantly reduce AI's ecological footprint without hindering its development and role in advancing society.

02

AI'S ENERGY CONSUMPTION

Artificial intelligence has become a major driver of energy consumption, with inference accounting for 60 to 80 percent of total energy use in many real-world applications. This demand has been propelled by the exponential growth of generative AI tools such as ChatGPT and their integration into global infrastructure.

The AI lifecycle consists of two key phases: training and inference. Training refers to the process of developing an AI model by exposing it to vast datasets to learn patterns, relationships, and decision-making rules. This phase involves iterative adjustments to the model's parameters, often numbering in the hundreds of billions, to optimize performance. Training is computationally intensive, requiring high-performance computing clusters equipped with thousands of GPUs or TPUs operating continuously for weeks or months. OpenAI's 175 billion-parameter model consumed 1,287 megawatt-hours during training, equivalent to the annual energy use of 130 U.S. households. Model size plays a critical role in training energy consumption, with DeepMind's 280 billion-parameter system using 1,066 megawatt-hours. The complexity of datasets also contributes significantly, as preprocessing petabytes of data, such as Common Crawl for GPT-3, introduces ancillary energy costs. Another driver of energy demand is hyperparameter tuning, where optimizing model configurations often requires multiple training runs, increasing energy consumption by 20 to 40 percent.

Between 2012 and 2018, training energy consumption grew at a tenfold annual rate, far outpacing efficiency improvements from Moore's Law. Innovations such as Google's Tensor Processing Units have helped mitigate this growth, enabling the company to train a model seven times larger than GPT-3 using 33 percent less energy. However, the relationship between model size and energy use is nonlinear, as doubling parameters can increase energy consumption by three to five times due to memory bandwidth bottlenecks and communication overhead in distributed training. While pre-training accounts for 90 to 95 percent of total lifecycle energy in large models, fine-tuning's cumulative impact grows with frequent, decentralized use by developers.

Inference, which occurs after deployment, is the process of using trained models to generate outputs in response to new inputs. While inference consumes far less energy per task than training, its cumulative impact is significant at scale. A single ChatGPT response requires 0.047 kilowatt-hours, comparable to streaming Netflix for three and a half minutes. At the scale of one trillion queries annually, that unit figure implies roughly 47 terawatt-hours (0.047 kilowatt-hours across a trillion queries is 47,000,000 megawatt-hours) — more than Ireland's entire national electricity consumption, and on the order of 36,000 times the energy used to train GPT-3 once. Two caveats belong with that number. It is an upper bound: the 0.047 kilowatt-hour figure sits at the high end of published estimates, which commonly range from about 0.002 to 0.005 kilowatt-hours per query, and every total below scales linearly with whichever value is used. And it is a projection from a per-query unit cost, not a measured system-wide total. Several factors amplify inference energy consumption, including the rapid adoption of generative AI, the expansion of edge computing, and the real-time demands of applications such as autonomous vehicles. Tools like DALL-E and Midjourney consume 2.9 kilowatt-hours per 1,000 image generations, which is 360 times more than text classification tasks. Deploying AI in billions of smartphones and IoT devices further complicates energy measurement. Real-time applications require continuous inference, preventing energy-saving idle states and adding to the overall power demand.

Leading technology companies report that inference now accounts for the majority of their AI-related energy budgets, surpassing training in operational impact. At scale, inference energy consumption can dwarf the energy used in training. Training GPT-3 required 1,287 megawatt-hours, but if it were to serve 4.3 billion users each submitting one query per day, annual inference energy consumption would reach roughly 74 terawatt-hours (4.3 billion users across 365 days at 0.047 kilowatt-hours a query is about 73,800,000 megawatt-hours) — on the order of 57,000 times the energy used to train the model once. The two scenarios are consistent with each other by construction: this one represents about 1.57 trillion queries against the one trillion above, and yields about 1.57 times the energy. Both are illustrative arithmetic on a single unit figure rather than measurements, and both should be read as upper bounds for the reason given above.

The environmental impact of AI is further amplified by the carbon intensity of the energy sources used to power the underlying infrastructure. For example, training GPT-3 on a coal-powered grid emits 552 tons of carbon dioxide, whereas using a renewable-powered grid reduces emissions to just 24 tons. However, it's important to note that many cloud data centers increasingly rely on partially renewable energy mixes. For instance, if a cloud provider operates with an average of 40% renewable energy, then AI applications hosted on that infrastructure inherit that carbon profile — a critical calibration point when comparing emissions across different energy grids. The emissions from inference also vary significantly based on where queries are processed. A ChatGPT query in Wyoming, where 95 percent of electricity comes from coal, generates approximately ten times the emissions of the same query processed in Quebec, where 94 percent of power is hydroelectric. Tools like WattTime's estimation models, which use real-time, location-specific data to calculate emissions based on marginal power sources, are increasingly essential for more accurately assessing AI's carbon footprint at a granular level.

The disparity in emissions across different energy grids underscores the importance of sustainable AI practices. Several strategies have emerged to reduce AI's energy footprint, including powering data centers with renewables, implementing dynamic voltage scaling to reduce GPU power use by 25 percent, and adopting model compression techniques such as knowledge distillation to create smaller models with 60 percent lower inference costs. Carbon-aware scheduling, which optimizes training times to coincide with high renewable energy availability, has been shown to cut emissions by 30 percent.

AI energy consumption is typically measured using three primary metrics: kilowatt-hours, GPU or TPU hours, and carbon emissions. Kilowatt-hours quantify direct energy use, with training OpenAI's GPT-3 model requiring 1,287 megawatt-hours — equivalent to the electricity use of more than 100 U.S. households in a year. GPU or TPU hours reflect the duration and intensity of hardware utilization. Training a 110 million-parameter BERT model for 80 hours on 64 GPUs required 1,507 kilowatt-hours, while inference tasks on an NVIDIA A100 GPU operate at 120 to 200 watts depending on workload. Carbon emissions provide insight into the environmental trade-offs of AI development. That same 1,507 kilowatt-hour BERT-base training run, executed on a coal-powered grid at roughly one kilogram of carbon dioxide per kilowatt-hour, corresponds to about 3,300 pounds of carbon dioxide, whereas fine-tuning the model produced 26 pounds of carbon dioxide per run. The widely quoted 626,000-pound figure comes from the same study (Strubell et al., 2019) but describes something much larger: a full neural architecture search with hyperparameter tuning, comprising thousands of training runs rather than one. Conflating the two overstates the cost of training a single model by more than two orders of magnitude.

These figures illustrate the growing impact of AI on global energy demand. Data centers worldwide already consume between 240 and 340 terawatt-hours annually, and as AI adoption accelerates, addressing energy efficiency and sustainability will become a critical priority.

DATA CENTERS & HARDWARE IMPACT

Data centers are the backbone of modern AI, providing the computational power needed to train and run machine learning models. These facilities house thousands of high-performance servers, processing vast amounts of data for cloud computing, AI applications, and digital services. AI data centers are energy-intensive, consuming an estimated 1–2% of global electricity, a figure projected to rise with AI becoming more advanced and popular. The source of this energy, either renewable or fossil fuels, significantly impacts their sustainability. Additionally, cooling systems, which are necessary to prevent overheating, consume enormous amounts of water and electricity, raising concerns about the environment and carbon footprints. Beyond energy consumption, the production of AI hardware such as GPUs, TPUs, and custom AI chips, has its own environmental toll. As AI companies scale their infrastructure, major tech firms like Google, Microsoft, and Amazon are investing in greener AI strategies, pushing for renewable-powered data centers, liquid cooling systems, and more efficient AI chips. As of 2023 data, there were an estimated 11,000 data centers globally (Minnix, 2025).

As mentioned, AI models require immense computing power to process vast amounts of data, train deep learning models, and perform real-time inference. Specialized hardware such as GPUs (Graphics Processing Units), TPUs (Tensor Processing Units), and other AI accelerators have become a driver in modern AI infrastructure. GPUs were originally designed for rendering graphics in gaming and visual applications, but have become essential for AI due to their ability to process many tasks simultaneously. The leading manufacturers of GPUs are Nvidia (A100, H100), AMD (Instinct MI300), and Intel (Gaudi AI chips). TPUs are custom AI chips developed by Google, specifically designed to accelerate deep learning computations. Unlike GPUs, TPUs are optimized for tensor operations, making them highly efficient for training and running large AI models. They consume less power per computation compared to GPUs but are mostly available within Google Cloud's ecosystem. These AI chips, and the magnets, batteries, and power systems that support them, rely on critical minerals such as lithium, cobalt, and nickel, alongside true rare-earth elements — the lanthanides plus scandium and yttrium — used in high-performance magnets and electronics. All of them require energy-intensive mining operations. The extraction of these minerals contributes to deforestation and habitat destruction in mining regions (e.g., the Democratic Republic of the Congo for cobalt, China for rare earth metals), toxic waste and pollution as refining these materials generates hazardous byproducts, and high carbon emissions from transportation and processing.

Data centers have been at the forefront of controversies due to their high use of energy, water usage for cooling systems and more. When it comes to direct energy consumption, it is important to focus on not just IT equipment (servers consuming immense energy amounts) but general energy used for other systems from lighting to monitoring, given the size of data centers.

This means that apart from IT equipment energy consumption focusing on servers, storage arrays and networking devices, individuals should also consider ancillary systems energy consumption which focuses on cooling systems, power distribution and backup systems (e.g. uninterruptible power supplies, power distribution units). The issue with this is that additional energy loss is expected and electrical conversion inefficiencies add to this. This is one of the issues of energy use by data centers, and with more data centers being built and operated, overall energy consumption is increasing. One key metric to consider for power usage is PUE – power usage effectiveness – which compares total facility energy consumption to the energy used by IT equipment. Because the numerator includes the denominator, PUE is bounded below by 1.0 and is unbounded above; a value under 1.0 is physically impossible. A PUE of 1.0 is the ideal floor, meaning all energy goes to IT equipment and none to cooling, power distribution, or other overhead, while real data centers typically run between about 1.1 and 2.0 (Gillis, 2022). A lower PUE value means better overall energy efficiency, and this metric can be used to compare energy efficiency of different data centers or companies. With certain cooling systems being implemented more (that use less energy), PUE can be lowered. The metric was developed by The Green Grid, a consortium focusing on improving data center energy efficiency (Digital Realty).

Apart from IT equipment energy consumption, a key aspect of energy use is cooling technologies that are implemented in data centers. The main types of system to consider are the following:

Air Cooling: This type of cooling relies on air conditioning, which circulates cooled air throughout the data center using tubes. The typical setup includes a raised floor design where cool air is delivered through perforated tiles in designated cold aisles, while hot air is exhausted from hot aisles. Hot/cold aisle containment helps improve efficiency by physically separating hot and cold air streams, reducing mixing and lowering energy consumption. The main advantage of air cooling is its familiarity, relatively low upfront costs, and easy integration into existing infrastructure. However, as computing densities increase, air cooling may struggle to remove heat effectively, which leads to higher use of energy (Schaap, 2024).

Liquid Cooling: Liquid cooling is a more efficient alternative, using two types of cooling: direct-to-chip cooling and immersion cooling. In the first example, water or fluid circulates through plates. This is particularly beneficial for high-performance computing environments. Immersion cooling takes it a step further by submerging server components or racks in a liquid that absorbs heat directly from all surfaces. This method reduces reliance on air-based cooling, operates more quietly, and helps maintain lower operating temperatures, extending hardware lifespan. The issue with this example is the higher costs associated with the needed equipment (Vertiv).

Free Cooling: This example uses cooler outside air to replace hot air near equipment, as opposed to reusing the same air over and over that is cooled. This significantly reduces energy consumption and lowers operational costs. However, its effectiveness depends on location and climate variations as it works best when outside air is cool. Additionally, humidity and dust control must be carefully managed to prevent equipment damage (Masters DC).

Evaporative Cooling: This type of cooling has three forms – direct, indirect and two-stage. Evaporative cooling relies on the natural process of water evaporation to cool down outside air that is then blown onto equipment. When hot air meets pads filled with water, liquid turns into gas, cooling the facility, and remaining water from the pads is recirculated to keep the process ongoing. The main environment for this type of cooling would be in more humid areas (Airsys, 2024).

Hybrid Approaches: Many modern data centers use hybrid cooling systems, combining multiple technologies for optimal performance. For example, liquid cooling may be used for high-density racks, while air or free cooling is applied to other areas. This depends on the capability of each data center and environment – e.g. the need for one type of cooling system over another. Overall, liquid cooling has become more popular, but it is important to note that hybrid approaches are still considered the 'future,' as data centers become more efficient and a combination of different cooling techniques is used to increase efficiency (Robb, 2024).

A metric used for water use and efficiency is WUE – water usage effectiveness. This value measures how efficiently a data center uses water in its cooling and operational processes. A lower WUE indicates better water efficiency, meaning less water is used per unit of computing power, while a higher WUE suggests increased water consumption, which may be unsustainable in regions facing water scarcity both in the US and in other countries with higher temperatures and less access to water. Data centers in water-stressed areas must optimize cooling methods to minimize freshwater usage. For example, evaporative cooling systems reduce electricity consumption but require significant water use. Companies track WUE as part of their ESG goals, though maintaining a low WUE can be particularly challenging in climates where cooling efficiency must be balanced with limited water availability — a challenge to consider in future designs of data centers and sustainability goals (Higgins, 2024).

Given the substantial energy and resource demands of AI infrastructure, the need for sustainable solutions is more urgent than ever.

04

MITIGATION STRATEGIES & SUSTAINABLE AI

As AI models like GPT-4 and DALL·E grow larger and more complex, companies and academic institutions are working to address the immense energy demands required to train and operate these systems. Major players such as Google, Microsoft, Amazon, Meta, OpenAI, and leading universities are implementing innovative solutions to reduce AI's carbon footprint and promote sustainable AI development.

How to read the corporate claims in this section. Almost everything that follows is drawn from the companies' own sustainability pages and engineering blogs. These are the parties with the strongest incentive to present their record favourably, and three distinctions should travel with every figure. First, a pledge is not an achievement: a target set for 2030 or 2040 describes an intention, and the only evidence that it is being met is independently reported progress against it. Second, most corporate renewable-energy claims are market-based: the company buys renewable energy certificates or signs power purchase agreements sufficient to match its annual consumption, which is not the same as the location-based question of whether clean electricity was flowing on the grid at the hour and place the computation actually ran. Third, offsetting is not the same as not emitting; the quality and permanence of offsets is itself contested, and an offset tonne and an unemitted tonne are not interchangeable. The section reports what each company says, and marks where the claim is a pledge, where it rests on market-based accounting, and where independent reporting complicates it.

Google has been developing several projects in AI sustainability, and has set a target of powering its data centers, which host massive AI operations, with 100% carbon-free energy by 2030 (Google Data Centers, 2024). That is a target rather than a description of current operations, and it is a demanding one precisely because Google states it in hourly terms: matching consumption with clean energy on every grid, every hour, rather than annually in aggregate. Google's own reporting places its global hourly carbon-free share materially below 100%, and the gap between the annual-matching claim and the hourly one is where most of the difficulty in this commitment sits. To improve operational efficiency, Google employs AI-driven cooling systems, which have reduced the energy needed for cooling its data centers by up to 40% (Google DeepMind, 2016). The company also continues to develop custom Tensor Processing Units (TPUs) that are optimized to handle AI workloads more efficiently, using less electricity than conventional hardware (Google Cloud, 2018). Through these efforts, Google is reducing both the operational and computational energy demands of its AI systems.

Microsoft is also taking significant action to curb the environmental impact of AI. The company has pledged to be carbon-negative by 2030, which means removing more carbon from the atmosphere than it emits, including emissions from AI workloads run on Azure, its cloud platform (The Microsoft Cloud, 2024). The pledge should be read against the company's own subsequent environmental reporting, which has shown total emissions rising rather than falling since it was made, driven largely by the construction and operation of new datacentre capacity. That is the central tension of this whole section in miniature: efficiency commitments made in good faith can be outrun by growth in the underlying demand. To achieve this, Microsoft is investing in advanced clean energy solutions, such as battery storage and nuclear energy, to sustainably power AI data centers (Microsoft Sustainability, 2023). Additionally, Microsoft is working on optimizing AI models using techniques like Neural Architecture Search and muTransfer, which help reduce the energy needed to train and run AI systems. These methods focus on building more efficient models from the start, allowing AI to achieve high performance while minimizing carbon emissions (Microsoft Research, 2022).

Amazon Web Services (AWS), one of the largest providers of AI cloud computing, sits behind two commitments that an earlier version of this report ran together. Amazon's Climate Pledge commits the company to net-zero carbon across its operations by 2040. Separately, Amazon reports having matched its annual electricity consumption with renewable sources, a milestone it announced ahead of its own schedule (DataCentre Magazine, 2024). A target set for 2040 cannot have been met in 2024, and the two claims should not be merged. The renewable-matching claim is also market-based, resting on power purchase agreements and renewable energy certificates purchased across a year, an accounting method that has been criticised for allowing consumption in one region and hour to be offset by generation in another. AWS is also using real-time energy monitoring and workload distribution to ensure that computing resources are allocated efficiently, avoiding unnecessary energy use (About Amazon, 2024). In addition to these operational efforts, AWS is working on making its hardware more sustainable by focusing on recyclable and environmentally friendly server materials (Amazon Web Services, 2024).

Meta is working to reduce the carbon footprint of its AI systems through emissions transparency and infrastructure efficiency. Training its LLaMA 3-70B model generated an estimated 2,290 metric tons of CO_2, which Meta reports as offset by its sustainability initiatives (Hugging Face, 2024). Offset is the operative word: the tonnes were emitted, and a purchased offset is a claim that an equivalent quantity was avoided or removed elsewhere. The two are treated as equivalent in most corporate reporting and are not equivalent in the atmosphere, since offset quality, permanence and additionality are all contested. The same qualification applies to the claim of net zero operational emissions since 2020, which is a statement about scope one and scope two emissions net of renewable-energy matching and offsets, and does not cover the emissions embodied in manufacturing the hardware. Meta is also building AI-optimized data centers with liquid cooling and high energy efficiency (Meta Sustainability, 2024). It is worth noting that Meta's publication of a per-model training figure at all is unusual, and is a genuine transparency gain over competitors who publish nothing.

OpenAI, the developer of models like GPT-4 and DALL·E, operates some of the most computationally intensive AI systems in the world. However, OpenAI has not publicly released any sustainability strategy, carbon footprint disclosures, or commitments related to reducing emissions from its AI systems.

Finally, academic institutions are playing a key role in advancing sustainable AI practices. Researchers at the University of Michigan, for example, have developed training optimization techniques that could cut AI energy consumption by up to 75% (Champion, 2023). These methods focus on reducing redundant training processes and finding smarter ways to achieve high performance with less data and fewer computational cycles. At MIT, researchers created the 'Once-for-All' network, which cuts emissions by enabling one model to generate many smaller submodels without retraining — reducing training emissions by up to 1,300 times (MIT News, 2020). UMass Amherst is heading a $12 million NSF-backed initiative to cut computing emissions through integrated improvements across hardware and software systems (UMass Amherst, 2024). In addition, Yale and partner institutions are developing carbon reporting standards for machine learning to help shrink emissions across the sector (Yale School of the Environment, 2024).

The peer-reviewed literature is the appropriate counterweight to all of this, and it should be read directly rather than through the outlets that summarise it. Three studies generated most of the figures that circulate in reports of this kind. Strubell, Ganesh and McCallum (2019) produced the training-emissions numbers, including the 626,000 pound architecture-search figure so widely misattributed to training a single model. Patterson and colleagues (2021) measured the emissions of large neural network training directly and argued that the choice of model, processor, datacentre and grid together account for differences of two orders of magnitude, which is the finding that makes carbon-aware scheduling worth doing at all. Luccioni, Viguier and Ligozat (2022) estimated the footprint of the BLOOM model across its whole lifecycle, including hardware manufacture and idle consumption, and found that counting only the training run understates the total substantially. Where a corporate claim and one of these studies disagree, the study is the better authority.

Across both private companies and academic research, there is a growing recognition of the need to reduce AI's environmental impact. From optimizing model architectures and improving data center operations to designing efficient hardware and using renewable energy, these organizations are taking concrete steps to ensure that AI's future is not only powerful but also sustainable. While these efforts represent meaningful progress, continued innovation, investment in clean energy, and greater transparency will be essential to align AI's growth with global climate goals. The honest summary of this section is that the direction of travel is right, the pledges are mostly unmet by construction because their deadlines have not arrived, and the emissions curve has so far been driven more by growth in demand than bent by efficiency.

05

REGULATORY FRAMEWORKS

The environmental sustainability of AI remains largely overlooked in both environmental law and technology regulations. UNEP has criticized the lack of policy focus on AI's ecological impact, noting that current efforts prioritize AI's trustworthiness rather than its environmental footprint. Hacker (2023) similarly highlights this gap and provides policy recommendations to address it. The European Union has taken steps to regulate AI through its 2024 AI Act, the world's first comprehensive AI law, which classifies AI systems by risk levels. While the European Parliament emphasized making AI systems environmentally friendly, this aspect was largely neglected in the final legislation. Earlier drafts included sustainability provisions such as sustainable AI development principles, energy logging requirements, and foundation model standards, but these were weakened during negotiations. The final Act relies on voluntary codes and limited disclosure requirements, with Article 40 mandating that general-purpose AI providers disclose estimated energy consumption. Additionally, the European Energy Efficiency Directive requires data centers exceeding 500 kW to report energy use, while Germany has imposed stricter regulations, requiring data centers with over 300 kW capacity to source 50% of their electricity from renewables by 2024, increasing to 100% by 2027.

In the United States, the Sustainable Data Centers Act, introduced in December 2024, aims to regulate the environmental footprint of data centers critical to AI operations by requiring alignment with state renewable energy targets and annual reporting of energy and water usage. On a global scale, effective AI regulation requires international cooperation, particularly concerning AI's resource-intensive supply chains and the mining of critical minerals. In February 2025, the AI Action Summit marked a turning point, with ecological sustainability taking center stage. This led to the formation of the Coalition for Environmentally Sustainable Artificial Intelligence, initiated by France, UNEP, and the International Telecommunication Union (ITU). With over 100 partners, including 11 countries, five international organizations, and 37 tech companies, the coalition aims to drive momentum toward sustainable AI practices. UNEP plans to publish a guide in 2025 to promote energy-efficient data centers based on international best practices. Additionally, AI is being leveraged for climate solutions, such as optimizing renewable energy, monitoring environmental changes, improving climate modeling, and enhancing disaster preparedness.

Several initiatives are actively advancing AI-driven climate solutions, underscoring the potential of AI in achieving sustainability goals. The WEKA Sustainable AI Initiative focuses on responsible AI applications that improve data efficiency while supporting sustainable development goals, including investments in reforestation projects to offset carbon emissions from data centers (WEKA, n.d.). The UN Climate Change's AI for Climate Action initiative explores AI's role in advancing climate-resilient and low-emissions development, particularly in developing countries, with the goal of delivering transformative climate action through AI-powered solutions (UNFCCC, n.d.). The Bezos Earth Fund's AI for Climate and Nature Initiative seeks to leverage AI to address climate change and protect natural ecosystems, including a $100 million grand challenge to unlock AI-driven solutions for climate and nature problems (Bezos Earth Fund, 2025).

While technological advancements can help reduce AI's environmental impact, policy frameworks and ethical considerations play a critical role in ensuring sustainable AI development. Effective governance, regulatory measures, and corporate responsibility initiatives are necessary to align AI's growth with long-term environmental goals.

06

ETHICAL & POLICY CONSIDERATIONS

Government policies and international agreements are increasingly shaping the intersection of AI and sustainability, focusing on reducing the environmental impact of AI while promoting ethical and energy-efficient innovations. Various national strategies and global frameworks aim to balance technological advancement with ecological responsibility.

At the national level, governments are introducing AI sustainability policies to regulate energy consumption and carbon emissions. For example, the European Union's AI Act includes provisions for transparency in AI energy use, while the U.S. Department of Energy funds research on AI-driven energy efficiency. China's Green AI Initiative encourages data centers to adopt renewable energy sources and optimize power usage. Additionally, governments worldwide are promoting sustainable AI practices through tax incentives, grants, and stricter regulations on high-energy-consuming AI training processes.

Agreements like the Paris Agreement on Climate Change indirectly influence AI sustainability by urging nations to reduce carbon footprints, including those from digital infrastructures. The OECD AI Principles, adopted by over 40 countries, emphasize energy-efficient AI development and responsible deployment. The United Nations' Sustainable Development Goals (SDGs) encourage AI innovation to support climate action, particularly in renewable energy forecasting and environmental monitoring. Furthermore, the G7 and G20 summits have discussed AI's role in sustainable development, highlighting the need for global cooperation on eco-friendly AI deployment.

Despite these efforts, challenges remain in enforcing energy-efficient AI practices, as AI models grow more complex and require increasing computational power. As governments, organizations, and researchers work toward mitigating AI's environmental costs, it is essential to consider the broader implications of these efforts. Looking ahead, continued research, policy evolution, and cross-sector collaboration will be key to balancing AI's benefits with its sustainability challenges.

07

CONCLUSION & FUTURE DIRECTIONS

AI's rapid advancement presents both unprecedented opportunities and significant environmental challenges. As its adoption continues to grow, ensuring that AI development aligns with sustainability goals is no longer an option, but a necessity (World Economic Forum, 2024). While AI's energy consumption, reliance on resource-intensive data centers, and electronic waste generation pose serious environmental risks, innovative solutions in algorithmic efficiency, green data centers, and sustainable hardware development offer promising pathways forward (Nature, 2024; MIT News, 2025).

The future of AI sustainability depends on a multi-faceted approach that integrates technological innovation, policy intervention, and corporate responsibility. Society must prioritize energy-efficient AI architectures (Capitol Technology University, n.d.), while policymakers should implement regulations that incentivize renewable energy use and enforce transparency in AI's environmental impact (SIAM News, n.d.). Companies leading AI development must take proactive measures by investing in carbon-aware computing, energy-efficient chips, and responsible e-waste management.

Moving forward, AI should not only minimize harm but also become a key player in solving global environmental challenges. Its potential to optimize renewable energy distribution, enhance climate modeling, and improve conservation efforts demonstrates that AI can be part of the solution rather than just a contributor to the problem (AI Speakers Agency, n.d.; World Economic Forum, 2025). However, achieving this balance requires urgent and collaborative action across industries, governments, and academia. The question is no longer whether AI can be sustainable, but rather how quickly we can ensure that it is.

References

SOURCES

01 About Amazon Staff. (2024, June 25). How AWS helps reduce carbon footprint of AI workloads. About Amazon.

02 Adams, H. S. (2024, August 5). Study reveals AWS cloud can slash carbon emissions by 99% compared to on-prem data centers. DataCentre Magazine.

03 AI Speakers Agency. (n.d.). How will AI enable a sustainable future?

04 AIMultiple. (2025). Top 10 sustainability AI applications & real-life examples in 2025.

05 Airsys. (2024). Evaporative cooling for data centers – Pros and cons.

06 Bezos Earth Fund. (n.d.). AI for climate and nature initiative.

07 Brightlio. (2025, April 19). 170 data center stats (April 2025).

08 Capitol Technology University. (n.d.). Moving towards a more sustainable future using AI.

09 Champion, Z. (2023, April 17). Optimization could cut the carbon footprint of AI training by up to 75%. University of Michigan News.

10 Digital Realty. (n.d.). What is power usage effectiveness (PUE)?

11 Earth.org. (2025). The real environmental impact of AI.

12 European Parliament. (2023, August 6). EU AI Act: First regulation on artificial intelligence.

13 Evans, R., & Gao, J. (2016, July 20). DeepMind AI reduces Google data centre cooling bill by 40%. DeepMind.

14 EY. (n.d.). AI and sustainability: Opportunities, challenges, and impact.

15 Forbes. (2024, March 26). Here's why data center cooling is the hottest innovation in the sector [By A. Schaap].

16 Gillis, A. S. (2022, April 29). What is PUE (power usage effectiveness)? TechTarget.

17 Google. (2024). Operating sustainably – Google Data Centers.

18 Greenly. (n.d.). How can artificial intelligence help tackle climate change?

19 Hacker, P. (2023, December 21). Sustainable AI regulation. Semantic Scholar.

20 Harvard Business Review. (2024, July 15). The uneven distribution of AI's environmental impacts.

21 Higgins, A. (2024, November 13). What is water usage effectiveness (WUE) in data centers? Equinix.

22 Hugging Face. (2024). Meta-Llama-3-70B.

23 Intel. (n.d.). Advancing toward more sustainable AI.

24 Kunze, J. (2024, October 4). Environmental impacts of artificial intelligence: Peril, promise, and policy. Society for Industrial and Applied Mathematics.

25 Laranjeira De Pereira, J. R. (2024, April 8). The EU AI Act and environmental protection: The case for a missed opportunity. Heinrich Böll Stiftung, Brussels Office.

26 Latham & Watkins. (2024, July 12). US environmental, social, and governance legal considerations for AI companies.

27 Luccioni, A. S., Viguier, S., & Ligozat, A.-L. (2022). Estimating the carbon footprint of BLOOM, a 176B parameter language model.

28 Masters DC. (n.d.). How data center cooling works and why it is brilliant.

29 Matheson, R. (2020, April 23). Reducing the carbon footprint of artificial intelligence. MIT News.

30 McQuate, S. (2023, July 27). Q&A: UW researcher discusses just how much energy ChatGPT uses. UW News.

31 Meta. (2024). Meta sustainability report 2024. Meta Sustainability.

32 Microsoft. (2023). Improving sustainability with AI. Microsoft Sustainability.

33 MIT News. (2025). Explained: Generative AI's environmental impact.

34 National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1).

35 National Institute of Standards and Technology. (2024). AI risk management framework: Generative AI profile (NIST AI 600-1).

36 Nature. (2024). Ecological footprints, carbon emissions, and energy transitions.

37 Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon emissions and large neural network training.

38 Publications Office of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council.

39 Robb, D. (2024, June 3). Hybrid cooling: The bridge to full liquid cooling in data centers. Data Center Knowledge.

40 Russinovich, M. (2024, September 12). Sustainable by design: Innovating for energy efficiency in AI, part 1. The Microsoft Cloud Blog.

41 SIAM News. (n.d.). Environmental impacts of artificial intelligence: Peril, promise, and policy.

42 Sitecore. (n.d.). Harnessing the power of AI for a sustainable future.

43 Soto, K. (2018, August 30). What makes TPUs fine-tuned for deep learning? Google Cloud Blog.

44 Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–3650.

45 U.S. Congress. (2024). Artificial Intelligence Environmental Impacts Act, 118th Congress (S. 3732).

46 U.S. Department of Energy. (2024a). AI for energy: Opportunities for a modern grid and clean energy economy.

47 U.S. Department of Energy. (2024b, April 29). DOE announces new actions to enhance America's global leadership in artificial intelligence.

48 UN Environment Programme. (2025, February 11). New coalition aims to put artificial intelligence on a more sustainable path.

49 UN Framework Convention on Climate Change. (n.d.). AI for climate action.

50 University of California, Irvine Libraries. (n.d.). Ethics - AI in research.

51 Venkatesan, V., & Karibandi, A. (2024, September 16). Embracing modernization with a sustainability focus. Amazon Web Services.

52 Vertiv. (n.d.). Liquid cooling options for data centers.

53 WEKA. (n.d.). Sustainable AI initiative.

54 Westbrook, J. (2024, May 23). UMass researchers awarded $12M to shrink society's carbon footprint with AI & other computer science tools. UMass Amherst.

55 World Economic Forum. (2024). AI will accelerate sustainability – but is no silver bullet.

56 World Economic Forum. (2025). AI's role in the climate transition and how it can drive growth.

57 Yale School of the Environment. (2024). New initiative focuses on reducing the carbon footprint of computer systems and AI.

58 Yale School of the Environment. (2024, November 13). Can we mitigate AI's environmental impacts?