Hong Kong International Journal of Research Studies, ISSN: 3078-4018
Volume 3, Issue 2, July-December, 2025
Available online at: https://octopuspublication.com/index.php/hkijrs

Balancing AI-Led Technological Advancements with Environmental Sustainability

Sarvesh Chandirani

Year 13 | Jumeirah College, Dubai, UAE

ABSTRACT

While AI has established itself as the cornerstone of modern technological innovations, its accelerating growth brings substantial challenges. Increasingly complex and large scale AI models and the associated computing infrastructures consume massive volume of electricity, water and minerals, creating environmental challenges in the process. This paper examines the role of AI as a contributor to environmental challenges and also as a potential source of solution to those challenges. This examination looks into optimization of energy consumption in data centers, AI based dynamic management of workload, renewable energy integration and progresses in the field of energy efficient computing hardware and algorithms. In this process, this paper also looks into the challenges and limitations such as tradeoffs between performance and sustainability, entry barriers in the form of high capital investment, limitations of hardware, and ethical issues related to transparency, inclusiveness and accountability. Policy aspects such as the lack of globally accepted regulatory framework performance metrics are examined to show the gaps in governance. Based on the findings, this paper concludes that AI based sustainable computing requires a combination of technological innovation, ethical considerations and global standardization of performance metrics and frameworks. Through responsible use of AI, the optimal balance between environmental sustainability goals and computational performance can be achieved.

Keywords: Artificial Intelligence, Sustainable Computing, Green AI, Data Centers, Energy Optimization, Ethics, Policy

INTRODUCTION

Artificial Intelligence (AI) is no longer confined to the domains of specialized research as its pervasiveness shapes economic activities, technological innovations, and scientific researches. However, the environmental costs associated with the accelerating growth of AI becomes an area of concern. Emergence of large AI models, especially generative systems like Gemini and GPT, has led to rapid increase in demand for computational resources to facilitate training and inference, and it leads to increasing energy consumption and carbon emissions (Liu & Yin, 2024). This accelerating demand for computational resources and energy is straining the global energy systems that raises concerns over its sustainability.

Data centers are an integral part of AI infrastructure and need huge amount of energy, consuming roughly 2 to 3 percent of the worldwide electricity consumption (Gadepally, 2025). Based on some projections this energy consumption could rise to over 21% and AI related energy demand will play a crucial role in it (Gadepally,2025). In 2023 the data centers in the US consumed almost 4.4% of the total domestic electricity and persistence of this trend could see this figure almost triple by 2028 (Electric Power Research Institute [EPRI], 2024). Environmental strain of these massive data centers goes beyond just energy needs and includes the use of water for cooling purposes and creation of electronic waste, factors that further complicate the sustainability concerns (Sidorkin, 2025).

While AI infrastructure raises these sustainability related concerns, AI also seems to offer means to reduce its negative impact on the environment. Google's DeepMind can be seen as an apt example with a demonstrated ability to reduce the cooling energy required by data centers by up to 40%. Accomplished through reinforcement learning algorithm, it show how AI can also help with optimization of operational efficiency (Luo et al., 2022). Similarly, employment of predictive models can help with integration of cloud infrastructure with renewable energy sources and that would make computing more environment friendly (Sapre, 2024). Thus, the concern of environmental sustainability is both aggravated and alleviated by AI.

This paper explores how the environmental strain caused by increasing computational needs of AI models intersects with its role in optimizing computational infrastructure to make it more environmentally sustainable. In this process, paper addresses the main question: How can AI keep growing and shaping other technological innovations while also contributing to environmental sustainability? In the process of answering this question, it looks at the ways to optimize energy consumption, potential challenges and trade-offs, and considerations at ethical and policy making levels. Through analysis of relevant case studies, technological progresses and existing regulatory framework, this paper aims to find ways that will create balance between technological advancements and environmental sustainability.

Research Background

Artificial Intelligence is no longer confined within a narrow domain of research. It has become a technology with transformative ramifications for other technologies, economies, and societies. Its increasingly wider application already spans across healthcare, finance, processing of natural languages, and development of autonomous systems, making it a paradigm defining innovation of this century (Wu et al., 2021). However, this tremendously transformative role of AI also comes with very serious environetnal concerns. The computational power needed for the training and deployment of large scale models, based on deep learning and generative systems, requires increasingly larger amount of energy and resources (Tabbakh et al., 2024).

The growing scale of computation

There is a directly proportionate link between the increasing size and complexity of AI models and their environmental impact. Highly sophisticated AI models like GPT-4 and other large language models have an ever increasing number of parameters, already running in billions, that require rigorous training and the use of hardware specifically designed for these processes. These energy intensive training processes need tens of gigawatt/hours of electricity. It means a single hour of training for such large scale models may be equal to the energy consumed by thousands of households over a year (Chen et al., 2025). AI workloads are different from traditional computing as they require high-performance computing infrastructure on a sustained basis to perform optimization across large datasets through continuous iterations.

Resource intensive Data Centers

This ever increasing scale of AI operations is built upon increasingly massive data centers that host not only the training clusters but also the interference service. The energy guzzling nature of these data centers can be estimated from the fact that in 2024 the global energy consumption by data centers was 415 terawatt-hours (TWh) of electricity that constitutes around 1.5 percent of the global demand. This figure is expected to rise above 3% by the year 2030 and AI workload will play a key role in it (Chen et al. , 2025). The electricity consumption by data centers in the USA accounted for almost 4.4% of total uses in 2023 and this figure is expected to rise above three times by 2028 with the persistence of ongoing trends (EPRI, 2024). These massive data centers require powerful cooling systems that make up around 40% of the total electricity consumed (Virta Ventures, 2024). The global distribution of data centers may see some region specific concentration leading to grid stress and higher carbon footprints, especially when electricity production uses fossil fuels.

Besides massive amount of electricity, these data centers also put stress over other critical resources like water and minerals. Hyperscale data centers often use water evaporation method in cooling process and that needs massive amount of fresh water. Use of such processes in regions that are already experiencing water scarcity could further worsen the situation and bring industrial needs in conflict with public interest (UNEP,2024). Moreover, the AI infrastructure requires specialized computing hardware like GPUs and TPUs that are capable of high performance but their manufacturing process involves rare earth minerals. These minerals are processed through semiconductor fabrication, which is also highly energy intensive. Thus, the computing infrastructure needed for AI models, both the data centers and the specialized computing hardware, place significant stress on energy and minerals and raise concerns over its environmental sustainability (Sharp, 2025).

The impact of AI models on the environment extends beyond the training phase. Once the AI model is trained, it is deployed to process user queries. These queries result in continuous consumption of energy, and while the energy consumption per inference process could be lower then training, the massive amount of queries makes the total energy consumption quite substantial over the period (Wu et al., 2021). Besides the energy needs, the specialized hardware needed for infrastructure also undergo periodic upgrades to handle the increasing computation demand, and it leads to substantial electronic waste.

AI as the solution

While the environmental concerns associated with AI have dominated the literature, the idea of ‘Green AI’ has started coming into vogue. Green AI focusses upon creating algorithms and hardware that are energy efficient and contribute to sustainable practices within data centers to maintain the computational performance but at much reduced environmental cost (Raman et al., 2024). The focus remains on techniques that require lesser computational power without compromising accuracy- model pruning, quantization, and knowledge distillation. Besides these techniques, specialized hardware like Tensor Processing Units and Field Programmable Gate Array offer better energy efficiency in terms of performance-per-watt in comparison to conventional Graphic Processing Units (Tabbakh et al., 2024).
Besides algorithmic and hardware based efficiency, AI models are also emerging as helpful tools for sustainability. AI based predictive models can be used to ensure optimal energy efficiency in the data centers and they can do it through efficient distribution of workload and adjustment in cooling system accordingly. Deepmind project by Google is a fine example of AI enabled energy efficiency as it reduced the energy consumption in cooling processes by up to 40%.

Based on algorithms using reinforcement learning, DeepMind sets an example of energy efficiency facilitated by AI (Evans & Gaon, 2016; Luo et al., 2022). The role of AI to achieve to energy efficiency and environmental sustainability goes beyond its applications in data centers as it improves integration of renewable energy through more accurate output forecasts, helping grid operators in more effective and efficient balance of demand and supply (Razak et al., 2025).

Opportunities and Benefits

Artificial intelligence has the potential to promote sustainable computing through increased efficiency in the use of energy and water, reduction in the emissions of greenhouse gases and enhancement of operational resilience of computing infrastructure. This potential opportunity offered by AI can be divided into four components, where all the components complement each other and provide a realistic pathway to maintain performance and value of AI while reducing its environmental footprint. These four components are:

A) **Optimization of energy consumption**- Among the computing infrastructure used by AI, data centers function as the backbone, and cooling the data centers is a key energy consuming process. Under normal deployment, it may consume almost 30 to 40% of the energy used by the facility and , hence, the best target for AI-based optimization (Virta Ventures, 2024). The DeepMind project by Google sets an inspiring example as it shows how reinforcement learning agents that have been extensively trained on live telemetry can autonomously adjust the chiller plants, cooling towers and setpoints to maintain the safety standards while minimizing energy consumption. Early findings showed reduction in cooling energy by up to 40 % (Evans & Gao, 2016). Further studies in real world applications generalize these findings. Deployment of safety aware reinforcement learning (RL) agents as controllers saves around 9 to 13% energy across commercial facilities, as per the documented findings by Google and Trane Technologies (Luo et al., 2022). These technological interventions improve the effectiveness of power usage and reduce indirect greenhouse gas emissions.

Besides optimization of energy consumption, AI-based cooling system can also reduce consumption of freshwater, particularly when combined with dynamic shift in modes among evaporative, adiabatic and liquid-cooling. RL based autonomous scheduling of setpoints when the ambient humidity is higher or there is water scarcity, and prevention of inefficient drifts through predictive maintenance are some ways to reduce the water footprint of these operations (UNEP, 2024, Zewe, 2025). Although retrofitting the hardware like direct-to-chip cooling require lots of capital, adding AI control demonstrates improvement across key metrics- performance-per-watt and liters-per-compute- in regions with water scarcity (Raman et al., 2024).

B) **Making efficient models and dynamic workload management** - This component is based on the principle of finding the best fit between model and timing of use. A key aspect of Green AI philosophy is adoption of algorithm and architecture that maintain accuracy while minimizing computation load (Tabbakh et al., 2024). This efficiency is through multi-pronged strategies:

i) One strategy is pruning, quantization, distillation. Pruning involves removal of parameters that are redundant. Quantization lowers precision like shift from FP16 to INT8. Distillation is the process in which knowledge is transferred from a large teacher model to a smaller teacher model. Combination of these steps can reduce the Floating-point Operations Per Seconds (FLOPs) for each inference and need for memory bandwidth, while also maintaining the performance level for the tasks (Tabbakh et al., 2024).

ii) Another strategy is the use of conditional computation where only a subset of experts is activated for each token, which reduces the average computing needed for each query. Optimization of routing provides better speed for each quantity, thus reducing the computational overhead and performance bottlenecks (Chen et al., 2025).

iii) Distinguishing the complexity level of tasks and intelligent deployment of models also helps with resource efficiency. In this strategy, simpler or routine tasks like basic Q&A, extraction or classification of data are routed to compact models while frontier systems are reserved for complex tasks. It reduces total energy consumption while maintaining the standard o user experience (Raman et al., 2024).

iv) Another strategy to reduce the compute load per query is to cache the frequent responses and use retrieval-augmented generation (RAG) to avoid generation steps that are inessential. This strategy is especially effective in enterprise settings where knowledge lookups are used repeatedly (OECD, 2024).

Besides these algorithmic and architectural strategies, AI-powered scheduling system uses predictive demand to shift the workloads based on suitable timing and region. For instance, a non-urgent batch of inference may be deferred to timings with higher availability of renewable energy or an inference task may be shifted to a region with lower grid stress. This automated shift in time and region reduces environmental footprint per compute and eases pressure on the local grids (EPRI, 2024). When this autonomous and intelligent shift based on time and space is combined with dynamic right-sizing- finding best fit between uncertainty or confidence threshold and model variant- there is optimal provision for resources in the system and energy consumption is further reduced (Raman et al., 2024).

Intelligent integration of renewable energy

The impact of AI extends beyond computational infrastructure to the power grids as well. Its ability to forecast wind and solar energy output levels in the short term helps with better commitment decisions, reduces the costs of balancing energy supplies, and raises the contribution of renewable energy in electricity supplied to data centers (Razak et al., 2025). DeepMind project by Google can predict output level of wind energy 36 hours in advance, and it improved value of wind energy through better day-ahead scheduling. It shows how AI improves the economy of renewable dispatch (Razak et al., 2025).

The grid system further benefits from AI through better energy storage dispatch and demand response. In campuses that are grid-connected, the coordination of battery charging/discharging for storage of energy during peak hours, absorption of surplus solar power, and provision for frequency support is handled by RL based controllers. In the massive data centers, this strategy lowers the costs of grid interconnection, reduces the demand charges and stimulates renewable energy Power Purchase Agreement through better alignment of compute and generation profiles (EPRI, 2024; Che et al., 2025). Predictive analytics, at a larger scale, guide the siting process based on availability of cleaner energy, lower temperature and ample water, thus improving the operational emissions and environmental impact of cooling (Nature News eature, 2025; OECD, 2024).

Governance, reporting and market signals

Measurement is a key requirement for achieving operational excellence. Reporting Power Usage Effectiveness and purchases of renewable energy credits are some common practices nowadays, but these reports do not provide effective measurement of true lifecycle impact that includes manufacturing emissions, water footprints, and generation of e-waste (OECD, 2024; UNEP, 2024). Standardization of AI-centric metrics like energy consumed per token, water footprint per query, carbon footprint per training run, and emissions liked with hardware can provide more effective benchmarking and also show the returns from different interventions (EPRI, 2024).

Once these metrics are in place, AI can automate accounting based on these metrics. Operators and regulators will have actionable data through the AI automated carbon and water footprint dashboard based on telemetry feed, detection of anomalies in outlier facilities, and attribution models that show emission associated with model training, fine-tuning, and inference (OECD, 2024). This increased transparency, facilitated by AI, can reduce greenwashing and lead to more genuine progress towards environmental sustainability.

Economic benefits and system resilience

Just sustainability may not be effective driving force for the businesses and, hence, further benefits need to be identified. Operational optimization through AI saves energy costs, reduces demand charges and defer capital expenditures through better utilization of resources (Luo et al., 2022). Use of AI improves thermal management, which extends the lifespan of hardware, reduces the volume of e-waste and emissions, and also cuts down the expenditure on replacing the equipment (Virta Ventures, 2024). AI also boosts the system resilience as the intelligent controls are more effective in maintaining uptime in adverse conditions, protecting service level agreements and reducing the chances of deviation from recommended temperature range (Nature News Features, 2025).

The RL based agents create data post deployment that further refines their control policies and forecasting models, making them increasingly efficient and provide more savings (Luo et al., 2022). When combined with strategies like task segregation based on importance or complexity, shifting workload, and renewable PPAs, the overall effect can lead to significant decrease in energy cost and emission per compute (EPRI, 2024).

Challenges and Limitations

While AI presents ways of achieving sustainable computing, its pursuit has inherent challenges and limitations. These challenges and limitations cover technical, economic and ethical aspects, adding to the complexity of ‘green AI’ endeavour.

One of the key challenges of ‘Green AI’ is finding optimal balance between performance and sustainability. Training and inference processes of complex AI models like large language models (LLMs) need billions of parameters and reducing energy consumption in these processes require techniques like pruning and quantization for model compression that may come at the expense of accuracy and functionality (Rman et al., 2024). It creates a major dilemma- whether to go for high quality performance or for resource-efficiency that may constrain innovation (Tabbakh et al., 2024). This trade off between performance and resource efficiency becomes quite critical in sectors where speed and accuracy are prioritized.

The computing infrastructure required for AI have heavy reliance on specialized hardware like GPUs and TPUs. These hardware are not only expensive but also energy-intensive. While newer processors are more energy efficient and provide better performance-per-watt ratio, manufacturing these processors is a complex process and requires rate earth elements that contributes to environmental degradation (UNEP,2024). Moreover, the upgradation of hardware to satisfy

the increasing demand for computation power also adds to e-waste and creates concerns for environmental sustainability (Sharp, 2025).

Implementation of AI based sustainability solutions often face entry barriers with very high initial investment. Upgrading the data centers by integrating cooling systems with AI controls or AI-based integration technologies for renewable energy needs substantial capital investment (Wu et al., 2021). It could make these innovations unaffordable for smaller organizations or developing economies, thus increasing the digital divide (Toderas, 2025). Besides this, these sustainable AI based initiatives can achieve intended return on investment only in the long terms and it would disincentivize organizations that are looking for short term profitability.

Energy optimization through AI systems relies on data produced through extensive networks of sensors. Although this operational data is integral to continuously improving predictive modeling, it also creates challenges related to privacy and cyber-security (Shaddick et al., 2025). The infrastructure-related data is sensitive in nature and any unauthorized access can lead to major security breach. It necessitates the presence of very strong framework for encryption and data compliance, making the implementation of sustainability measures more complex.

There is no global standard to account for the environmental impact of AI, which further undermines the pursuit of sustainable computing. While there are regional initiatives- for example the EU Green Deal favours carbon-neutral technologies- there is lack of consensus among global frameworks (UNEP,2024). In the absence of global standards for accounting, organizations may find it difficult to benchmark the performance of their sustainability measures (Razak et al., 2025).

Another limiting aspect of AI driven sustainability is its assumption that AI is singlehandedly capable of overcoming sustainability challenges. This assumption may overlook the need for systemic changes for long term solutions like taking steps towards overall reduction in computing demands, and encouraging consumption of computational power in more responsible manner (Tabbakh et al., 2024).

Ethical and Policy consideration

Using AI to achieve sustainable computing presents challenges at the level of ethics and policy making. While there are documented evidences of how it can make computing more efficient and environmentally sustainable, there remain concerns regarding fairness, accountability and governance.

Sustainability oriented AI systems rely on massive datasets for predictive modeling of consumption and optimization of operations. Any biases present in these datasets may lead to undesirable outcomes. For example, a biased data set may result in biased algorithms prioritizing energy savings in developed regions that have better infrastructure at the expense of developing regions with poorer infrastructure, thus increasing inequalities at the global level (Shaddick et al., 2025). There should be due consideration for establishing data collection systems that offer better transparency and inclusive design that accommodate geographical and scio-economic diversity (Raman et al., 2024).

A major concern about transparency is that the decision-making processes of AI models remain opaque. While RL models are used to optimize cooling processes or scheduling of workload, their operational aspects remain largely inaccessible for stakeholders and they could not comprehend and assess the decision-making process (Wu et al., 2021). Since there is no transparency in operational processes, determining accountability in the event of an error like energy mismanagement or falling behind sustainability goals becomes quite difficult. Hence, there is need for explainable AI frameworks that provide outputs that can be interpreted and audited (Razak et al., 2025).

AI based sustainability initiatives are often marketed by organizations as their CRS programs. Since there are no globally standardized metrics for measuring the outcome of these initiatives, their sustainability benefits may be overstated to gain better market reputation, presenting the risk of ‘greenwashing’ (UNEP, 2024). It requires the presence of effective ethical governance where sustainability claims are subject to independent verification and organizations are bound to disclose carbon footprint of their AI operations (Tabbakh et al., 2024).
The lack of a universal framework to govern the environmental impact of AI is a major concern. European Union has taken the initiatives of Green Deal and a proposed AI act that place emphasis upon carbon neutrality and ethical practices, many other regions in the world lack similar frameworks (UNEP, 2024). Similarly, the lack of globally standardized metrics to measure energy efficiency and environmental footprint makes the enforcement and compliance inconsistent across the regions. There is a need for global agreement over a framework that balances computational innovations with environmental sustainability (Toderas, 2025).

With the increasing autonomy of AI systems, ethical concerns pertaining to control and responsibility will acquire greater significance. Placing the entire responsibility of operational efficiency on AI systems without human oversight is a big question. Although there are documented evidence of AI systems bringing operational efficiency and leading towards sustainable computing, the reduced human agency in the management of critical infrastructure is a big risk

(Shaddick et al., 2025). It is imperative for the policymakers to clearly define the boundaries of Autonomy for AI and take measures that align the sustainability goals with human and democratic values.

CONCLUSION

Artificial Intelligence(AI) presents a very unique problem in the discourses about sustainability. On one hand, it increases environmental stress and on the other hand it also offers possible solutions for them. With its increasing role in modern technological innovations, AI technology has undergone rapid growth, resulting in accelerating energy demand, especially the huge data centers that sustain the complex models. This increasing demand for AI technology underlines the urgency needed in aligning technological innovations with environmental sustainability, thus making computing practices sustainable.

This paper explored the ways AI can mitigate its environmental footprint through optimized energy uses, RL based management of workload, and integration of renewable energy. DeepMind Project of Google reducing energy consumption of cooling process by almost 40% demonstrates how AI based sustainability measures can produce measurable outcomes. Moreover, combination of progress in energy efficient hardware (better GPUs and TPUs) with energy efficient algorithms (through pruning and quantization) has proven successful in providing technical solutions that can find the optimal balance between resource consumption and performance.

However, there are significant challenges in the pursuit of using AI for sustainable computing. There is a tradeoff between computing performance and environmental sustainability. High capital investment and the environmental costs associated with manufacturing of hardware further complicate this pursuit. There is also a need for strong framework of governance to address ethical concerns like biases in algorithms, lack of transparency, and accountability. In the absence of globally standardized metrics and regulatory framework, sustainability claims are at risk of becoming unreliable or misleading.

A multipronged approach is needed to achieve sustainable computing through AI. The technical solutions offered by AI need complementary interventions at the level of global policymaking in the form of standardized metrics and frameworks.

REFERENCES

[1]. Chen, X., Wang,X., Colacelli, A., Lee, M., & Xie, L. (2025). Electricity demand and grid impacts of AI data centers: Challenges and prospects. arXiv
[2]. Electrical Power Research Institute [EPRI]. (2024). Powering intelligence: Analyzing artificial intelligence and data enter energy consumption.
[3]. Evans, R., & Gao, J. (2016). DeepMind AI reduces Google data center cooling bill by 40%. Google DeepMind Blog.
[4]. Luo, J., et al. (2022). Controlling commercial cooling systems using reinforcement learning. arXiv preprint arXiv: 2211.07357
[5]. Nature News Features. (2025, March 5). How much energy will AI really consume? The good, the bad and the unknown.
[6]. OECD. (2024). Measuring the environmental impacts of artificial intelligence compute and applications.
[7]. Raman, R., Pattnaik, D., Lathabai, H.H., et al. (2024). Green and sustainable AI research: An integrated thematic and topic modeling analysis. Journal of Big Data, 11(55)
[8]. Razak, T.R., Ismail, M.H., Darus, M.Y., Jarimi, H., & Su, Y. (2025). Artificial intelligence in renewable energy: a systematic review of trends in solar, wind, and smart grid applications. Research and Reviews in Sustainability , 1(1), 1-22.
[9]. Shaddick, G., Topping,D., Hales, T.C., et al. (2025). Data science and AI or sustainable futures: Opportunities and challenges. Sustainability , 17 (5), 2019
[10]. Sharp, A. (2025). The environmental impact of artificial intelligence. The Organization for World Peace.
[11]. SiliconANGLE. (2025, October 16). Google's DeepMind and CFS are building an AI plasma control system for nuclear fusion.
[12]. Tabbakh, A., Al Amin, L., Islam, M., et al. (2024). Towards sustainable AI: A comprehensive framework for Green AI. Discover Sustainability, 5(408).
[13]. Toderas, M. (2025). Artificial intelligence for sustainability: A systematic review and critical analysis. Sustainability, 17(17), 8049.
[14]. UNEP. (2024). AI has an environmental problem. Here's what the world can do about that. United Nations Environment Programme.
[15]. Virta Ventures. (2024). AI data centers – operating an energy-efficient data center. Virta Ventures Insights.
[16]. Wu, C.J., Raghavendra, R., Gupta, U., et al. (2021). Sustainable AI: Environmental implications, challenges and opportunities. arXiv preprint arXiv:2111.00364.