

Towards Environmentally Equitable AI

MOHAMMAD HAJIESMAILI*, University of Massachusetts, Amherst, United States

SHAOLEI REN, University of California, Riverside, United States

RAMESH K. SITARAMAN, University of Massachusetts, Amherst, United States

ADAM WIERMAN, California Institute of Technology, United States

The skyrocketing demand for artificial intelligence (AI) has created an enormous appetite for globally deployed power-hungry servers. As a result, the environmental footprint of AI systems has come under increasing scrutiny. More crucially, the current way that we exploit AI workloads' flexibility and manage AI systems can lead to wildly different environmental impacts across locations, increasingly raising environmental inequity concerns and creating unintended sociotechnical consequences. In this paper, we advocate environmental equity as a priority for the management of future AI systems, advancing the boundaries of existing resource management for sustainable AI and also adding a unique dimension to AI fairness. Concretely, we uncover the potential of equity-aware geographical load balancing to fairly re-distribute the environmental cost across different regions, followed by algorithmic challenges. We conclude by discussing a few future directions to exploit the full potential of system management approaches to mitigate AI's environmental inequity.

1 INTRODUCTION

The growing adoption of artificial intelligence (AI) has been accelerating across all parts of society, boosting productivity and addressing pressing global challenges such as climate change. Nonetheless, the technological advancement of AI relies on computationally intensive calculations and thus has led to a surge in resource usage and energy consumption. Even putting aside the environmental toll of server manufacturing and supply chains, AI systems can create a huge environmental cost to communities and regions where they are deployed, including air/thermal pollution due to fossil fuel-based electricity generation and further stressed water resources due to AI's staggering water footprint [12, 25]. To make AI more environmentally friendly and ensure that its overall impacts on climate change are positive, recent studies have pursued multi-faceted approaches, including efficient training and inference [5], energy-efficient GPU and accelerator designs [19], carbon forecasting [14], carbon-aware task scheduling [1, 21], green cloud infrastructures [2], sustainable AI policies [10, 18], and more. Additionally, data center operators have also increasingly adopted carbon-free energy (such as solar and wind power) and climate-conscious cooling systems, lowering carbon footprint and direct water consumption [8].

Although these initiatives are encouraging, unfortunately, a worrisome outcome — *environmental inequity* — has emerged [3]. That is, minimizing the total environmental cost of a globally deployed AI system across multiple regions does not necessarily mean that each region is treated equitably. In fact, the environmental cost of AI is often disproportionately higher in certain disadvantaged regions than in others. Even worse, AI's environmental inequity can be amplified by existing environmental equity *agnostic* resource allocation, load balancing, and scheduling algorithms and compounded by enduring socioeconomic disparities between regions. For example, geographical load balancing (GLB) algorithms that aggressively exploit regional differences to seek lower electricity prices and/or more renewables [7, 17] may schedule more workloads to water-inefficient

*All the authors contributed equally and are listed in alphabetical order of last name.

Authors' addresses: Mohammad Hajiesmaili, University of Massachusetts, Amherst, United States; Shaolei Ren, University of California, Riverside, United States; Ramesh K. Sitaraman, University of Massachusetts, Amherst, United States; Adam Wierman, California Institute of Technology, United States.

data centers (located in, for example, water-stressed Arizona), resulting in a disproportionately high water footprint and adding further pressures to local water supplies [9].

Addressing the emerging environmental inequity is becoming an integral part of responsible AI [3]. It has increasingly received public attention and urgent calls for mitigation efforts. For example, the AI Now Institute compares the uneven regional distribution of AI’s environmental costs to “historical practices of settler colonialism and racial capitalism” in its 2023 Landscape report [11]; the United Nations Educational, Scientific and Cultural Organization (UNESCO) recommends against the usage of AI if it creates “disproportionate negative impacts on the environment” [23]; California recognizes the need for “ensuring environmental costs are equitably distributed” in its State Report [4]; and environmental justice is ranked by Meta as the most critical factor among all environmental-related topics [15].

In this paper, we advocate *environmental equity* as a priority for the management of future globally deployed AI systems. Concretely, we explore the potential of harnessing AI workloads’ scheduling flexibility and utilizing equity-aware GLB as a lever to fairly re-distribute the environmental cost across regions, ensuring that no single region disproportionately bears the environmental burden. Then, we present key algorithmic challenges to enable AI’s environmental equity without significantly degrading the other performance metrics, such as the energy cost and inference accuracy. Finally, we discuss future directions to unleash the full potential of system management for environmentally equitable AI, including coordinated scheduling of AI training and inference, joint optimization of IT and non-IT resources, holistic control of system knobs, and building theoretical foundations.

Our proposal of environmental equity advances the boundaries of existing research on sustainable AI and mitigates the otherwise uneven distribution of AI’s environmental costs across different regions. Additionally, equity and fairness are crucial considerations for AI. The existing research in this space has predominantly tackled prediction unfairness against disadvantaged individuals and/or groups [20, 26]. Thus, environmental equity adds a unique dimension of fairness and significantly complements the existing literature, collaboratively building equitable and responsible AI.

2 OPPORTUNITIES AND CHALLENGES FOR EQUITY-AWARE GLB

In this section, we present the potential opportunities of leveraging equity-aware GLB to fairly re-distribute the environmental cost across different regions, followed by algorithmic challenges.

2.1 Opportunities

The limited power grid capacity has necessitated increasing flexibility from data centers to support demand response and maintain grid stability. A notable example is the recent industry initiative to maximize load flexibility for grid-integrated data centers [6]. Specifically, AI workloads exhibit three primary types of flexibility: (1) *Spatial*: AI training and inference tasks can be distributed across multiple data centers with minimal impact on latency. (2) *Temporal*: AI training tasks can be executed intermittently, provided they meet a given deadline. (3) *Performance*: A single inference request can be processed by different AI models, each offering distinct trade-offs between accuracy and resource consumption. These flexibilities can be exploited to promote environmental equity while satisfying other performance objectives. To achieve this, we can leverage a variety of approaches, such as AI computing resource allocation, load balancing and job scheduling, which we collectively refer to as system *knobs*.

In practice, the data center fleet of large companies such as Google and Microsoft often includes a few tens of self-managed hyperscale data centers and many more leased third-party colocation data center spaces spreading throughout the world [8]. By renting virtual machines on public

clouds, even a small business can flexibly choose its deployment region and place its computing workloads accordingly. As such, GLB is an important and common knob that can spatially balance computing workloads' energy demand as well as environmental footprint across different locations.

As a concrete example, we consider moving AI inference workloads around from one data center to another and exploit equity-aware GLB to mitigate AI's environmental inequity. To achieve equitable distribution of AI's environmental cost, we consider the notion of *minimax* fairness. Mathematically, denoting $x_{i,t}$ as the amount of AI workloads processed in data center i at time t and $E_{i,t}(x_{i,t})$ as the resulting regional environmental cost (e.g., due to water consumption [12] and air/thermal/waste pollution from non-renewable energy [24]), we consider an equity-aware objective: $\sum_{t=1}^T \sum_i \text{cost}_{i,t}(x_{i,t}) + \lambda \cdot \max_i [\sum_{t=1}^T E_{i,t}(x_{i,t})]$, where the first term is the traditional GLB cost (e.g., total carbon/water footprint and energy cost) specified based on the prior literature [9], the second term " $\max_i [\sum_{t=1}^T E_{i,t}(x_{i,t})]$ " serves as the equity regularizer by reducing the highest regional environmental cost, and $\lambda \geq 0$ is the weight.

A snapshot of results. We run a simulation based on the BLOOM model (a large language model) inference trace deployed in 10 different data centers throughout the world and show a snapshot of our results in Table 1. The details of the

GLB	Metric	Algorithm				
		GLB-Cost	GLB-Carbon	GLB-Dist	eGLB-Off	eGLB
Full	Cost (US\$)	29170	45535	47038	33669	33752
	PAR (Water)	1.71	1.85	1.44	1.27	1.37
	PAR (Carbon)	1.68	1.70	1.41	1.13	1.22
Partial	Cost (US\$)	29659	45535	47038	34186	34162
	PAR (Water)	1.72	1.84	1.44	1.30	1.38
	PAR (Carbon)	1.69	1.71	1.41	1.12	1.22

Fig. 1. Comparison of GLB algorithms in terms of the total energy cost and the normalized water/carbon peak-to-average ratio (PAR). Details in [13].

simulation are available in [13]. We consider both *full* GLB (i.e., each request can be flexibly routed to any data center) and *partial* GLB (i.e., each request can only be routed to a subset of data centers depending on its originating location). Compared to common baseline algorithms that simply minimize the total energy cost (GLB-Cost), carbon emission (GLB-Carbon) or workload-to-data center distance (GLB-Dist), our algorithm (called eGLB-Off) can effectively mitigate the environmental inequity by reducing the ratio of the maximum to the average regional environmental footprint. Importantly, while there is an inevitable conflict between minimizing the total cost/environmental footprint and addressing the environmental inequity, eGLB-Off can still keep the total cost reasonably low. Additionally, we study a simple online algorithm (called eGLB) based on dual mirror descent to show the potential of mitigating environmental inequity in an online setting. While there is a gap between eGLB and eGLB-Off due to online informational constraints, eGLB outperforms the equity-unaware baseline algorithms in terms of the environmental footprint's peak-to-average ratio, demonstrating the potential of online GLB to mitigate AI's environmental inequity.

2.2 Challenges

While equity-aware GLB can potentially mitigate AI's environmental inequity, the equity regularizer " $\max_i [\sum_{t=1}^T E_{i,t}(x_{i,t})]$ " fundamentally separates our problem from the existing sustainable GLB approaches and creates substantial algorithmic challenges. Specifically, the equity cost " $\max_i [\sum_{t=1}^T E_{i,t}(x_{i,t})]$ " is unknown until the end of T time slots, but complete future information (e.g., future workload arrivals and water efficiency) may not be perfectly known in advance. Moreover, even though prediction is often available in practice, it may not be accurate, and its untrusted nature means we cannot simply take the prediction as if it were the ground truth.

Additionally, the traditional design of online competitive algorithms often focuses on guaranteeing the worst-case performance robustness. But, the resulting average performance can be far

from optimal due to the conservativeness needed to address potentially worst instances. By contrast, machine learning (ML) based optimizers, e.g., reinforcement learning policies, can improve the average performance of online decision-making by exploiting rich historical data and statistical information, but they typically sacrifice the strong performance robustness needed by real AI systems, especially when there is a distributional shift, the ML model capacity is limited, and/or inputs are adversarial. Thus, in order to achieve the best of both worlds while pursuing online equitable-aware GLB, we have to carefully balance the usage of traditional competitive algorithms and ML-based optimizers by designing new learning-augmented online algorithms.

3 FUTURE DIRECTIONS

We discuss a few future directions to leverage system knobs for environmentally equitable AI.

Coordinated scheduling of AI training and inference. While AI inference offers spatial flexibility, AI model training has great *temporal* scheduling flexibility as we can choose *when* to train the AI models in a stop-and-go manner. We can also choose where to perform AI model training and even possibly change the locations in the middle of the training process. Thus, a potential direction is to explore coordinated scheduling of AI training and inference tasks to fairly distribute AI's overall environmental costs across different regions.

Joint optimization of IT and non-IT resources. Data centers have increasingly begun to install on-site carbon-free energy, such as solar power, to partially power the workloads and lower the environmental footprint [21]. However, renewables are often intermittent, and the available energy storage capacity is finite. Thus, how to optimize AI demand response given intermittent renewables is challenging, yet worth investigating for addressing AI's environmental inequity.

Holistic control of system knobs. In addition to GLB, a rich set of system knobs are available and offer flexible tradeoffs, such as dynamic model selection for inference, turning servers on/off, and resource allocation to different AI tasks. For example, different AI models can exhibit different energy-accuracy tradeoffs for the same task. Holistic control of these system knobs holds enormous potential to curb AI's resource usage and mitigate environmental inequity, but also presents additional challenges due to the significantly enlarged decision space.

Theoretical foundations. Optimizing a variety of system knobs for environmentally equitable AI has its roots in *fair* decision-making, which is a classical area that bridges computer systems and algorithms and enjoys a long history with rich theoretical results [16, 22] and prominent production deployments. However, this classic literature primarily focuses on algorithms that ensure that different job or flow types receive a fair share of system resources, e.g., CPU, memory, etc. Tackling the challenges raised by environmental inequity in modern planet-scale AI systems requires a revisit to the algorithmic foundations and the development of new theoretical tools, which can systematically capture the conflicts between traditional measures of performance, such as accuracy and latency, with measures of emerging importance, such as environmental equity. Thus, it is crucial to build new theoretical foundations to support the design of environmentally equitable AI.

4 CONCLUSION

In light of AI's wildly different environmental costs across different regions, we advocate *environmental equity* as a priority for the management of future AI systems. We present the potential opportunities and algorithmic challenges of tapping into AI workloads' scheduling flexibility and leveraging equity-aware GLB to mitigate AI's environmental inequity. Finally, we discuss a few future directions to unleash the full potential of system knobs for environmentally equitable AI, including coordinated scheduling of AI training and inference, joint optimization of IT and non-IT resources, holistic control of system knobs, and building theoretical foundations. Our proposal of

environmental equity pushes forward the boundaries of existing system management for sustainable AI and also adds a unique dimension to AI fairness, collaboratively building equitable and responsible AI.

ACKNOWLEDGEMENT

The work of Mohammad Hajiesmaili is supported by NSF CNS-2325956, CAREER-2045641, CPS-2136199, CNS-2102963, CNS-2106299, and NGSDI-2105494. The work of Shaolei Ren is supported by NSF CCF-2324916. The work of Ramesh K. Sitaraman is supported by NSF CNS-2325956, NGSDI-2105494, and CNS-1763617. The work of Adam Wierman is supported by NSF CCF-2326609, CNS-2146814, CPS-2136197, CNS-2106403, and NGSDI-2105648.

REFERENCES

- [1] Bilge Acun, Benjamin Lee, Fiodar Kazhamiaka, Kiwan Maeng, Udit Gupta, Manoj Chakkaravarthy, David Brooks, and Carole-Jean Wu. 2023. Carbon Explorer: A Holistic Framework for Designing Carbon Aware Datacenters. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2* (Vancouver, BC, Canada) (ASPLOS 2023). Association for Computing Machinery, New York, NY, USA, 118–132. <https://doi.org/10.1145/3575693.3575754>
- [2] Noman Bashir, Tian Guo, Mohammad Hajiesmaili, David Irwin, Prashant Shenoy, Ramesh Sitaraman, Abel Souza, and Adam Wierman. 2021. Enabling Sustainable Clouds: The Case for Virtualizing the Energy System. In *Proceedings of the ACM Symposium on Cloud Computing* (Seattle, WA, USA) (SoCC '21). Association for Computing Machinery, New York, NY, USA, 350–358. <https://doi.org/10.1145/3472883.3487009>
- [3] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) (FAccT '21). Association for Computing Machinery, New York, NY, USA, 610–623. <https://doi.org/10.1145/3442188.3445922>
- [4] California Government Operations Agency. 2023. Benefits and Risks of Generative Artificial Intelligence Report. *State of California Report* (November 2023).
- [5] Lingjiao Chen, Matei Zaharia, and James Zou. 2023. FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance. arXiv:2305.05176 [cs.LG]
- [6] EPRI. 2024. DCFlex Initiative. <https://msites.epri.com/dcflex>.
- [7] Peter Xiang Gao, Andrew R. Curtis, Bernard Wong, and Srinivasan Keshav. 2012. It’s not easy being green. *SIGCOMM Comput. Commun. Rev.* (2012).
- [8] Google. 2023. Environmental Report. <https://sustainability.google/reports/>.
- [9] Mohammad A. Islam, Kishwar Ahmed, Hong Xu, Nguyen H. Tran, Gang Quan, and Shaolei Ren. 2018. Exploiting Spatio-Temporal Diversity for Water Saving in Geo-Distributed Data Centers. *IEEE Transactions on Cloud Computing* 6, 3 (2018), 734–746. <https://doi.org/10.1109/TCC.2016.2535201>
- [10] ISO/IEC JTC for AI (SC42). 2023. ISO/IEC TR 20226 Sustainability: Harnessing the Power of AI. <https://etech.iec.ch/issue/2023-06/sustainability-harnessing-the-power-of-ai>.
- [11] Amba Kak and Sarah Myers West. 2023. AI Now 2023 Landscape: Confronting Tech Power. *AI Now Institute* (April 2023).
- [12] Pengfei Li, Jianyi Yang, Mohammad A. Islam, and Shaolei Ren. 2024 (accepted). Making AI Less “Thirsty”: Uncovering and Addressing the Secret Water Footprint of AI Models. *Commun. ACM* (2024 (accepted)).
- [13] Pengfei Li, Jianyi Yang, Adam Wierman, and Shaolei Ren. 2024. Towards Environmentally Equitable AI via Geographical Load Balancing. In *e-Energy*.
- [14] Diptyaroop Maji, Prashant Shenoy, and Ramesh K Sitaraman. 2022. CarbonCast: Multi-day Forecasting of Grid Carbon Intensity. In *Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. 198–207.
- [15] Meta. 2021. Sustainability Report. <https://sustainability.fb.com/>.
- [16] Jeonghoon Mo and Jean Walrand. 2000. Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on Networking* 8, 5 (2000), 556–567.
- [17] Jorge Murillo, Walid A Hanafy, David Irwin, Ramesh Sitaraman, and Prashant Shenoy. 2024. CDN-Shifter: Leveraging Spatial Workload Shifting to Decarbonize Content Delivery Networks. In *Proceedings of the 2024 ACM Symposium on Cloud Computing*. 505–521.
- [18] OECD. 2022. Measuring the Environmental Impacts of Artificial Intelligence Compute and Applications: The AI Footprint. *OECD Digital Economy Papers* 341 (2022). <https://doi.org/https://doi.org/10.1787/7babf571-en>

- [19] David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluís-Miquel Munguia, Daniel Rothchild, David R. So, Maud Texier, and Jeff Dean. 2022. The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. *Computer* 55, 7 (2022), 18–28. <https://doi.org/10.1109/MC.2022.3148714>
- [20] Dana Pessach and Erez Shmueli. 2022. A Review on Fairness in Machine Learning. *ACM Comput. Surv.* 55, 3, Article 51 (February 2022), 44 pages. <https://doi.org/10.1145/3494672>
- [21] Ana Radovanović, Ross Koningstein, Ian Schneider, Bokan Chen, Alexandre Duarte, Binz Roy, Diyue Xiao, Maya Haridasan, Patrick Hung, Nick Care, Saurav Talukdar, Eric Mullen, Kendal Smith, MariEllen Cottman, and Walfredo Cirne. 2023. Carbon-Aware Computing for Datacenters. *IEEE Transactions on Power Systems* 38, 2 (2023), 1270–1280. <https://doi.org/10.1109/TPWRS.2022.3173250>
- [22] Rayadurgam Srikant and Lei Ying. 2013. *Communication networks: an optimization, control, and stochastic networks perspective*. Cambridge University Press.
- [23] UNESCO. 2022. Recommendation on the Ethics of Artificial Intelligence. In *Policy Recommendation*.
- [24] U.S. EPA. [n.d.]. About the U.S. Electricity System and its Impact on the Environment. <https://www.epa.gov/energy/about-us-electricity-system-and-its-impact-environment>.
- [25] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, et al. 2022. Sustainable AI: Environmental Implications, Challenges and Opportunities. In *Proceedings of Machine Learning and Systems*, Vol. 4. 795–813.
- [26] Xueru Zhang and Mingyan Liu. 2021. *Fairness in Learning-Based Sequential Decision Algorithms: A Survey*. Springer International Publishing, Cham, 525–555. https://doi.org/10.1007/978-3-030-60990-0_18