Towards Socially and Environmentally Responsible AI

Pengfei Li[*]
UC Riverside

Yejia Liu[*]
UC Riverside

Jianyi Yang
UC Riverside

Shaolei Ren[†]
UC Riverside

Abstract

The sharply increasing sizes of artificial intelligence (AI) models come with significant energy consumption and environmental footprints, which can disproportionately impact certain (often marginalized) regions and hence create environmental inequity concerns. Moreover, concerns with social inequity have also emerged, as AI computing resources may not be equitably distributed across the globe and users from certain disadvantaged regions with severe resource constraints can consistently experience inferior model performance. Importantly, the inequity concerns that encompass both social and environmental dimensions still remain unexplored and have increasingly hindered responsible AI. In this paper, we leverage the spatial flexibility of AI inference workloads and propose equitable geographical load balancing (GLB) to fairly balance AI's regional social and environmental costs. Concretely, to penalize the disproportionately high social and environmental costs for equity, we introduce L_q norms as novel regularization terms into the optimization objective for GLB decisions. Our empirical results based on real-world AI inference traces demonstrate that while the existing GLB algorithms result in disproportionately large social and environmental costs in certain regions, our proposed equitable GLB can fairly balance AI's negative social and environmental costs across all the regions.

1 Introduction

In the rapidly evolving field of artificial intelligence (AI), a significant transformation is underway with the emergence of large foundation models as exemplified by Large Language Models (LLMs) like GPTs [4] and Vision Transformers Models (ViTs) [8]. These cutting-edge AI models demonstrate the ability to function effectively in diverse contexts, engaging with extensive vocabularies and image data for unforeseen AI tasks, i.e., zero-shot abilities. To serve inference requests, they are typically deployed across geographically distributed data centers for better service availability, lower transmission latency, and/or privacy regulations.

Environmental inequity. Powerful yet hungry, large AI models require substantial resources not only during training but also in deployment and inference. For some popular AI services such as text and image generation, the total energy consumption for inference can be comparable to or even exceed that for training, resulting in huge carbon emissions and freshwater usage [14, 32]. To curb the growing environment footprint, many recent efforts have been devoted to enhancing the efficiency and reducing the energy consumption of AI models. Example strategies include model compression that reduces AI's computational demand for inference (typically at a sacrifice of model performance) [17, 18] and geographical load balancing (GLB) that leverages spatial heterogeneities to route more workloads to low-cost and/or greener regions [6, 15]. Additionally, on the infrastructure side, there has been a rise in the adoption of carbon-free energy and climate-conscious cooling system designs in the data center industry. For instance, utilizing air-side economizers where climate conditions allow has become increasingly common to cut the direct water consumption [21].

While these approaches can effectively minimize AI's total environmental footprint, the rise of *environmental inequity* — AI's negative environmental impact can disproportionately affects certain (often marginalized) regions [2, 15] — has become increasingly worrisome, potentially leading to other unintended social and ecological consequences and widening regional disparities. Importantly, the disproportional distribution of AI's environmental cost across different regions can be amplified by existing approaches to managing AI systems (e.g., load distribution and AI model scaling) that often prioritize the total environmental cost rather than the cost borne by individual regions which are most environmentally vulnerable [15]. Compounded by the sharply growing demand, AI's environmental inequity has received calls for mitigation efforts from various organizations, such as UNESCO [27], Meta [20] and the State of California [5].

Social inequity. Going beyond environmental footprints, concerns with AI's social inequity have also emerged [28]. For now, only a few major tech players have the resource and capacity to train and deploy large AI models. Thus, due to the uneven deployment of computing resources across the globe, users from different regions may encounter varying AI model sizes and performances (e.g., larger AI models typically imply better inference performance in terms of the accuracy and task scores), leading to complex societal consequences. For example, studies have indicated that people are becoming increasingly reliant on LLMs for acquiring knowledge, suggesting that subpar LLMs could jeopardize the prospects of these individuals [25].Thus, AI's potentially unfair model performance has close relevance to its social

[*] Equal contribution

[†] Corresponding author: Shaolei Ren (shaolei@ucr.edu)

inequity. Crucially, the existing environmentally conscious approaches to AI system management (e.g., choosing larger AI models with better performances/accuracies when there are more solar energy available) may further reinforce AI's performance unfairness among users from different regions, enlarging the social inequity.

Contributions. With the growing need for AI as a public resource serving the broader society, it becomes increasingly imperative to rectify AI's emerging social and environmental inequities and enable truly responsible AI [20, 24]. In this paper, we focus on the AI inference stage and introduce a novel equity-aware GLB algorithm to fairly balance AI's social and environmental costs across different regions. More specifically, we consider the performance cost of heterogeneous AI models and the carbon and water footprints associated with AI model inference by dynamically scheduling users' inference requests (a.k.a. workloads) using GLB. When optimizing GLB decisions, we leverage L_q norms in terms of AI's social and environmental costs as regularization terms to penalize decisions that disproportionately affect certain regions. In other words, regions with higher environmental and/or social costs will be prioritized and given a larger weight when leveraging GLB to minimize the total cost. By doing so, both the social and environmental costs of AI inference are more evenly distributed across different regions, thus mitigating AI's social and environmental inequities.

To assess the effectiveness of our method on promoting socially and environmentally equitable AI, we conduct a simulation-based case study of 10 geographically-distributed data centers serving inference requests for an LLM over an 18-day period. Our empirical results demonstrate that while the existing GLB algorithms result in disproportionately large social and/or environmental costs in certain regions, our proposed equitable GLB can fairly balance AI's negative social and environmental costs across all the regions.

2 Related Works

From the social fairness perspective, much attention has been directed towards protecting groups with certain attributes [16, 23, 30]. The issue is partially rooted in inherent biases within datasets and could potentially be exacerbated by models [16, 30]. To address such unfairness, numerous strategies have been developed. For instance, [3, 19] suggest removing sensitive attributes from datasets to prevent the model from relying on them, while others adjust prediction outcomes after training [22, 23]. Additionally, some have advocated for equivalent metrics, such as error rates, among specific groups [1, 7]. These studies typically focus on the model training stage, but the attained fairness can be compromised if AI models of different sizes are not equitably chosen for users from different regions. By stark contrast, we focus on the AI inference stage and judiciously balance the user requests from different regions across geographically distributed data centers hosting heterogenous AI models.

To address AI's environmental impacts, existing studies primarily focus on minimizing environmental metrics such as the total carbon emission, water footprint, or a weighted combination thereof, to enable environmentally responsible AI model training and inference [6, 14, 32]. Nonetheless, concerns with AI's environmental inequity across different regions have remained largely unaddressed. A recent study [15] has proposed to tackle the uneven distribution of AI's regional environmental costs via GLB. But, this approach overlooks the social equity dimension, which is equally, if not more, important element of responsible AI.

3 Problem Formulation

We focus on the AI inference stage and consider a set of pre-trained AI models denoted by $\mathcal{K} = \{1, 2, \cdots, K\}$, each with different performance and energy consumption for serving an inference request. There are a set of geographically distributed data centers $\mathcal{N} = \{1, 2, \cdots, N\}$ serving users coming from a set of regions $\mathcal{J} = \{1, 2, \cdots, J\}$.

Operational cost. At each time t, data center i dynamically selects one or more of the available heterogeneous AI models to serve the incoming workloads. More formally, we denote $y_{i,j}^k(t) \geq 0$ as the workload dispatched from region j to data center i served through model k at time t. Given the scheduled demand $y_{i,j}^k(t)$, we denote the energy consumption and computational resources necessary for deploying model k in data center i as $e_{i,k}(y_{i,j}^k(t))$ and $r_{i,k}(y_{i,j}^k(t))$, respectively. For example, both $e_{i,k}(y_{i,j}^k(t))$ and $r_{i,k}(y_{i,j}^k(t))$ can be modeled as linearly increasing functions in terms of $y_{i,j}^k(t)$. Thus, the total energy consumption at data center i can then be calculated as

$$e_i(t) = \sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} e_{i,k}(y_{i,j}^k(t)).$$

For notational simplicity, we define the set of workload distribution decisions at time t as $y(t) = \{y_{i,j}^k(t) | i \in \mathcal{N}, j \in \mathcal{J}, k \in \mathcal{K}\}$. We also take the energy price $p_{i,t}$ and power usage effectiveness (PUE, which accounts for non-IT energy overheads) γ_i of data center i into consideration. As a result, the total operational cost at time t can be written as

$$cost_t(y(t)) = \sum_{i \in \mathcal{N}} \gamma_i \cdot p_{i,t} \cdot \left[\sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} e_{i,k}(y_{i,j}^k(t)) \right]. \quad (1)$$

Social inequity cost. We define the noramlized performance cost of the AI model k as $s_k(y_{i,j}^k(t)) = s_k \cdot y_{i,j}^k(t) \geq 0$, where $s_k \geq 0$ represents the inference performance degradation cost for each request when using model k compared to the best possible model (usually the largest model [26]). For example, when model l has the best performance, its performance cost is zero for any allocated request. Here, the performance cost can be measured in terms of various metrics of an AI model (e.g., average inference accuracy and score of an LLM for a set of target tasks, among others). Thus, the total performance cost of the workload from region j is computed as $\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}} s_k(y^k_{i,j}(t))$, which, when normalized by the total workload $\lambda_{j,t}$, represents the AI model's average social performance for users from region j (i.e., a type of *group* fairness [23]). To balance AI's performance for users from different regions, we introduce a social fairness function $f_t(y(t))$ in terms of L_q norm of the average performance costs for users from different regions:

$$f_t(y(t)) = \left[\sum_{j\in\mathcal{J}}\left[\frac{\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}} s_k(y^k_{i,j}(t))}{\lambda_{j,t}}\right]^q\right]^{1/q}, \qquad (2)$$

where $q \geq 1$ is a hyperparameter that promotes AI's social equity for users from different regions. Concretely, we only care about the average AI model performance across different regions when $q = 1$ (i.e., no consideration of AI's social equity), whereas we focus on minimizing AI's worst regional model performance when $q \longrightarrow \infty$ (i.e., solely considering AI model performance for users from the most disadvantaged regions). The priorities for these two conflicting objectives are adjusted by varying $q \geq 1$.

Environmental inequity cost. Carbon emissions associated with fossil fuels and water consumption are the two main non-negligible factors. Besides the global warming effects, carbon emissions have significant local effects such as high air pollution and even elevated immortality rates [13], thus making it necessary to balance AI's regional carbon emissions. Depending on the fuel mix for electricity generation, the carbon emission rate can vary significantly across different physical locations and times of the day. Specifically, the carbon emission of data center i is denoted as $c_{i,t}(e_i(t))$, where $e_i(t)$ is the total energy consumption for running AI inference in data center i at time t. In general, an increased proportion of carbon-intensive energy sources (e.g. hard coals) directly correlates with higher carbon emissions, impacting the function $c_{i,t}(\cdot)$. The water consumption of deploying AI models is another important environmental cost and can be divided into two categories: onsite and offsite [14]. For each data center, onsite water is evaporated to reject the heat generated by servers into the outside environment (if the data center uses cooling towers), or cool and humidify the air entering the data center (if the data center uses air-side free cooling) [14]. The offset water refers to the water consumed for the electricity generation. In total, we define the water consumption as $w_{i,t}(e_i(t))$, which considers both onsite and offsite water and is linearly increasing with $e_i(t)$ depending on the runtime water usage effectiveness.

The total environmental cost of data center i is defined as

$$\mathcal{H}_i\left(\sum_{t=1}^{T} y(t)\right) = \sum_{t=1}^{T}\Big[\mu_c c_{i,t}\big(e_i(t)\big) + \mu_w w_{i,t}\big(e_i(t)\big)\Big],$$

where the hyperparameters $\mu_w \geq 0$ and $\mu_c \geq 0$ convert the carbon emission and water consumption to a single unit cost and balance their relative importance. By applying the L_q norm, the overall environmental inequity cost is defined as

$$g\left(\sum_{t=1}^{T} y(t)\right) = \left[\sum_{i\in\mathcal{N}}\left(\mathcal{H}_i\left(\sum_{t=1}^{T} y(t)\right)\right)^q\right]^{\frac{1}{q}}, \qquad (3)$$

where $q \geq 1$ prioritizes the minimization of AI's environmental cost in more disadvantaged data center locations/regions. In particular, when $q \longrightarrow \infty$, (3) becomes AI's worst environmental impact over all the data center locations.

GLB objective. We formulate the optimization objective of our socially and environmentally equitable GLB (called SE-GLB) as follows:

$$\min_{y(t),t=1,\cdots,T} \sum_{t=1}^{T} cost_t(y(t)) + \sum_{t=1}^{T} f_t(y(t)) + g\left(\sum_{t=1}^{T} y(t)\right) + \sum_{t=1}^{T}\sum_{i\in\mathcal{N},j\in\mathcal{J},k\in\mathcal{K}} y^k_{i,j}(t)\cdot d_{ij}, \qquad (4a)$$

$$s.t. \quad \sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}} y^k_{i,j}(t) = \lambda_{j,t}, \forall\ j \in \mathcal{J}, t = 1,\cdots,T, \qquad (4b)$$

$$\sum_{k\in\mathcal{K}} r_{i,k}\left(\sum_{j\in\mathcal{J}} y^k_{i,j}(t)\right) \leq M_i, \forall\ i \in \mathcal{N}, t = 1,\cdots,T \qquad (4c)$$

In (4a), the term $\sum_{i\in\mathcal{N},j\in\mathcal{J},k\in\mathcal{K}} y^k_{i,j}(t)\cdot d_{ij}$ accounts for the total moving cost for scheduling user requests from region j to data center i, where d_{ij} represents the moving cost for scheduling one unit of request (e.g., in proportion to the distance between region j and data center i). The constraint (4b) means that we need to schedule all the user demand $\lambda_{j,t}$ for each region j without request dropping, and the constraint (4c) denotes the computational resource constraint for AI inference in each data center i. Note that we can also easily add other constraints such as workload routing constraints (i.e., user requests from region j can only be routed to certain data center locations due to data sovereignty regulations or latency constraints).

Compared to the existing literature on GLB that typically minimizes the total cost or focuses on the environmental impact [9, 15], the key novelty of our formulation is to holistically address AI's social and environmental inequities by using L_q norms to penalize GLB decisions that lead to disproportionately high social and/or environmental costs in certain disadvantaged regions.

4 A Case Study

We run a simulation study to preliminarily validate SE-GLB to mitigate AI's social and environmental inequities.

4.1 Setup

We consider 10 geographically distributed data centers: four in the U.S. (Virginia, Georgia, Texas, and Nevada), four in Europe (Belgium, the Netherlands, Germany, and Denmark), and two in Asia (Singapore and Japan). Each of these locations hosts a large number of data centers. We also consider 10 regions, each corresponding to one distinct data center location in our experiments. To highlight the potential of equity-aware GLB, we consider *full* GLB flexibility, where workloads can be dispatched from any region to any data center. To host an LLM inference service, each data center contains a cluster of 150 identical Nvidia DGX A100 servers each equipped with eight NVIDIA A100 GPUs and a maximum power of 6.5kW. Excluding other services beyond our scope, each data center has a maximum AI inference server power of $\sim$ 1 MW. The data center PUE is set as 1.1 to adhere to efficient operation standards. The regional environmental impact is assessed using a weighted combination of carbon and water footprints. For inference, we assume three LLMs of different sizes are available: Llama-2-7B, Llama-2-13B, and Llama-2-70B [26].

Datasets. We utilize the GPU power trace spanning 18 days as used in [15]. We gather evaluation scores of Llama-2 from HuggingFace [31] across the model sizes of 7*B*, 13*B*, and 70*B* on benchmarks AI2 Reasoning Challenge, HellaSwag, and Truthful QA.We then average and normalize these scores for measuring AI's performance costs. Hourly energy prices across the 10 data centers are obtained from [10] for Europe and Asia, and from their respective ISOs for U.S. data centers [29]. Hourly weather data from [11] is utilized to calculate wet bulb temperature from dry bulb temperature and relative humidity. On-site WUE is determined using an empirical formula from [12].

Evaluation metrics. We consider four metrics: 1) *average energy cost*, calculated as the total energy cost over 18 days divided by the number of data center locations; 2) *average environmental footprint and social cost*, representing the total carbon emission, water footprint, and performance cost by the number of data center locations; 3) *maximum regional environmental footprint and performance cost*, which identifies the highest environmental and performance costs among the 10 data center locations and user regions; 4) *max/avg ratio*, representing the ratio of the maximum cost to the average cost for relative comparison. A lower value on this metric indicates a more equitable solution.

Baselines. 1) `Cost-GLB`: This algorithm optimizes the average energy cost and the performance cost. It can also be seen as a special case of `SE-GLB` where μ_c and μ_w are set as zero and $q = 1$. 2) `All-GLB`: This algorithm minimizes the weighted sum of the energy cost, environmental cost and societal cost (i.e., $q = 1$) based on [12]. 3) `E-GLB`: The environmentally equitable GLB algorithm which is studied in [15] and does not address AI's social inequity. Note that we do not consider the baseline that solely minimizes energy cost, as this approach would simply force the data centers to always choose the smallest model for inference.

Table 1. Comparison between different GLB algorithms.

Metric		**Algorithm**			
		`Cost-GLB`	`All-GLB`	`E-GLB`	**`SE-GLB`**
Energy (US$)	avg	83524	92945	101106	108197
Water (m^3)	avg	476.72	465.44	433.24	456.58
	max	1410.92	842.44	652.72	649.41
	max/avg	2.96	1.81	1.51	1.42
Carbon (ton)	avg	36.720	32.090	29.548	32.163
	max	110.275	55.491	41.923	48.054
	max/avg	3.00	1.73	1.42	1.49
Normalized Performance Cost	avg	0.262	0.244	0.268	0.248
	max	0.449	0.353	0.313	0.253
	max/avg	1.71	1.45	1.17	1.02
Performance Score	avg	57.83	58.11	57.74	58.05
	min	54.94	56.43	57.04	57.98

4.2 Results

We run an offline optimizer with all the future information in our case study, while online algorithms that optimize GLB without knowing future information are left as our future work. In Table 1, we show the cost comparison between baselines and `SE-GLB`. By default, the weights assigned to carbon emission and water consumption are $\mu_w = 60$ (US$/$m^3$) and $\mu_c = 1500$ (US$/ton), unless otherwise specified. We can observe that `Cost-GLB` has the lowest energy cost compared to other GLB approaches since it prioritizes the energy cost minimization. However, it also leads to the highest average carbon emission and water consumption. Additionally, `Cost-GLB` exhibits the highest max to average ratio in terms of social and environmental equities. Therefore, solely optimizing for energy cost can overburden certain regions with excessive workloads and worsen the AI's inequity. By comparison, `All-GLB` takes a weighted sum of energy cost and the environmental footprint, which reduces average water consumption and carbon emission. `E-GLB` further reduces the max to average ratio of environmental footprint by minimizing the L_q norm of AI's environmental impact across different locations. `SE-GLB` explicitly considers the L_q norms of both social and environmental costs, thereby achieving a more equitable distribution of AI's model performance and environmental impact across different user regions and data center locations. While this comes at an increased energy cost due to the conflict between equity and energy cost minimization, we argue that the cost increase is acceptable in order to mitigate AI's inequity that would otherwise create unintended socio-ecological consequences as people increasingly rely on AI.

5 Concluding Remarks

In this work, we holistically consider AI's social and environmental equity and propose novel equity-aware GLB to balance AI's regional social and environmental costs towards responsible AI. Our key novelty is to introduce L_q norms to penalize GLB decisions that would otherwise lead to disproportionately high social and/or environmental costs in disadvantaged regions. Our empirical evaluation has shown the effectiveness of our proposed approach in improving both social and environmental equity by prioritizing the most disadvantageous data centers and user regions.

Acknowledgement

This work was supported in part by the U.S. NSF under grants CNS-1910208, CNS-2007115, and CCF-2324941.

References

[1] Sina Baharlouei, Maher Nouiehed, Ahmad Beirami, and Meisam Razaviyayn. Rényi Fair Inference. *arXiv 1906.12005*, 2020.

[2] Rachel Bergmann and Sonja Solomun. From Tech to Justice: A Call for Environmental Justice in AI. *AI Now Institute*, October 2021.

[3] Sumon Biswas and Hridesh Rajan. Fair Preprocessing: Towards Understanding Compositional Fairness of Data Transformers in Machine Learning Pipeline. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE '21. ACM, August 2021.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. *arXiv 2005.14165*, 2020.

[5] California Government Operations Agency. Benefits and Risks of Generative Artificial Intelligence Report. *State of California Report*, November 2023.

[6] Andrew A Chien, Liuzixuan Lin, Hai Nguyen, Varsha Rao, Tristan Sharma, and Rajini Wijayawardana. Reducing the Carbon Impact of Generative AI Inference (Today and in 2035). In *Proceedings of the 2nd Workshop on Sustainable Computer Systems*, HotCarbon '23, New York, NY, USA, 2023. Association for Computing Machinery.

[7] Andrew Cotter, Heinrich Jiang, Maya Gupta, Serena Wang, Taman Narayan, Seungil You, and Karthik Sridharan. Optimization with Non-differentiable Constraints with Applications to Fairness, Recall, Churn, and Other Goals. *Journal of Machine Learning Research*, 20(172):1–59, 2019.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv 2010.11929*, 2021.

[9] Peter Xiang Gao, Andrew R Curtis, Bernard Wong, and Srinivasan Keshav. It's Not Easy Being Green. *ACM SIGCOMM Computer Communication Review*, 42(4):211–222, 2012.

[10] International Energy Agency (IEA). Data and Statistics. https://www.iea.org/data-and-statistics.

[11] Iowa State University. Iowa Environmental Mesonet. https://mesonet.agron.iastate.edu/.

[12] Mohammad A Islam, Kishwar Ahmed, Hong Xu, Nguyen H Tran, Gang Quan, and Shaolei Ren. Exploiting Spatio-temporal Diversity for Water Saving in Geo-distributed Data Centers. *IEEE Transactions on Cloud Computing*, 6(3):734–746, 2016.

[13] Mark Z. Jacobson. Enhancement of Local Air Pollution by Urban CO2 Domes. *Environmental Science & Technology*, 44(7):2497–2502, 2010. PMID: 20218542.

[14] Pengfei Li, Jianyi Yang, Mohammad A. Islam, and Shaolei Ren. Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models. *arXiv 2304.03271*, 2023.

[15] Pengfei Li, Jianyi Yang, Adam Wierman, and Shaolei Ren. Towards Environmentally Equitable AI via Geographical Load Balancing. In *e-Energy*, 2024.

[16] Tianlin Li, Qing Guo, Aishan Liu, Mengnan Du, Zhiming Li, and Yang Liu. FAIRER: Fairness as Decision Rationale Alignment, 2023.

[17] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration. *arXiv preprint arXiv:2306.00978*, 2023.

[18] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the Structural Pruning of Large Language Models. *Advances in neural information processing systems*, 36, 2024.

[19] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning Adversarially Fair and Transferable Representations. *arXiv 1802.06309*, 2018.

[20] Meta. Sustainability Report. https://sustainability.fb.com/, 2021.

[21] Meta Sustainability – Water. https://sustainability.fb.com/water/, 2023.

[22] Alejandro Noriega-Campero, Michiel A. Bakker, Bernardo Garcia-Bulle, and Alex Pentland. Active Fairness in Algorithmic Decision Making. *arXiv 1810.00031*, 2018.

[23] Dana Pessach and Erez Shmueli. A Review on Fairness in Machine Learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

[24] Shahana Rayhan. Ethical Implications of Creating AGI: Impact on Human Society, Privacy, and Power Dynamics. *Artificial Intelligence Review*, 2023.

[25] Kassym-Jomart Tokayev. Ethical Implications of Large Language Models A Multidimensional Exploration of Societal, Economic, and Technical Concerns. *International Journal of Social Analytics*, 8(9):17–33, Sep. 2023.

[26] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open Foundation and Fine-tuned Chat Models. *arXiv preprint arXiv:2307.09288*, 2023.

[27] UNESCO. Recommendation on the Ethics of Artificial Intelligence. In *Policy Recommendation*, 2022.

[28] United Nations General Assembly. Seizing the Opportunities of Safe, Secure and Trustworthy Artificial Intelligence Systems for Sustainable Development. Agenda item 13, Distr.: Limited, March 2024. Seventy-eighth session.

[29] U.S. Energy Information Administration. Open data. https://www.eia.gov/opendata/.

[30] Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. In-processing Modeling Techniques for Machine Learning Fairness: A Survey. *ACM Transactions on Knowledge Discovery from Data*, 17(3):1–27, 2023.

[31] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv 1910.03771*, 2020.

[32] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, et al. Sustainable AI: Environmental Implications, Challenges and Opportunities. In *Proceedings of Machine Learning and Systems*, volume 4, pages 795–813, 2022.