CHEMICAL ENGINEERING TRANSACTIONS
VOL. 107, 2023
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.cetjournal.it
Guest Editors: Petar S. Varbanov, Bohong Wang, Petro Kapustenko

ISBN 979-12-81206-07-6; **ISSN** 2283-9216
DOI: 10.3303/CET23107018

Sustainability of Large AI Models: Balancing Environmental and Social Impact with Technology and Regulations

Peter Szarmes, Gábor Élő*

Széchenyi István University, Egyetem tér 1., 9026 Győr, Hungary
elo@sze.hu

Artificial Intelligence (AI) systems, particularly large language models, have shown remarkable advancements, revolutionising various fields across industries. However, the sustainability of building large AI models with billions of parameters has become a subject of concern due to their significant environmental and social impact. The training of such models consumes enormous amounts of water and energy and emits substantial carbon emissions, contributing to climate change as data centres heavily rely on fossil fuels. This article summarises the current situation and explores the benefits and challenges of large AI models, emphasising the environmental impact and proposing strategies towards sustainability. Special attention is given to the social challenges, including accessibility, job displacement, biases, and data privacy concerns. Finally, the article advocates for the formulation of green and good AI practices standards for the future. To achieve sustainability, regulations are suggested to ensure transparency and accountability while promoting innovation-friendly frameworks. The authors see that while there is more progress in technology and infrastructure to address environmental impacts, social impacts are more neglected, and they are arguing for more detailed regulation as a solution.

1. Introduction

AI systems can improve climate models, develop materials for more energy-efficient infrastructure, manage renewable energy sources, detect environmental issues like methane leaks, monitor natural disasters, and optimise various systems for increased efficiency. Recently, large language models have become increasingly sophisticated and capable. These models, such as Microsoft's AI-powered Bing search engine, Google's Bard, and OpenAI's GPT-4, are trained on massive amounts of data to generate human-like text and perform various tasks previously thought impossible for computers.

Training a large language model involves exposing it to extensive datasets, which can take weeks or months. These models have millions or even billions of parameters that represent the relationships between words and are used for predictions. The training process requires powerful high-performance chips, often GPUs, which consume substantial energy. Then, the operation of servers and chips, as well as cooling systems, contributes to the substantial energy requirements. This energy consumption of AI systems is a major concern because data centres, where AI processing often occurs, rely heavily on fossil fuels, accounting for 2.5 to 3.7 % of global greenhouse gas emissions (Cho, 2023).

The next sections will demonstrate in more detail the benefits, the costs and the current challenges in this domain. The last sections show that regulations could help to overcome most of the challenges and suggest a general framework for more detailed objectives.

2. Benefits, costs and challenges of large AI models today

Despite the environmental and other concerns, large AI models offer several significant advantages. These models have demonstrated exceptional capabilities in natural language processing, computer vision, and other complex tasks, revolutionising various fields. These AI models have the potential to drive progress in application areas, including agriculture, biology, climate change, healthcare, scientific discovery, and many more. In healthcare, AI models can assist in disease diagnosis, drug discovery, and personalised medicine, improving

Paper Received: 23 June 2023; Revised: 25 July 2023; Accepted: 26 October 2023
Please cite this article as: Szarmes P., Élő G., 2023, Sustainability of Large AI Models: Balancing Environmental and Social Impact with Technology and Regulations, Chemical Engineering Transactions, 107, 103-108 DOI:10.3303/CET23107018

patient care and outcomes. For instance, researchers have developed AI models that can detect diseases with high accuracy. In climate modelling, large AI models can contribute to a better understanding and prediction of climate patterns, aiding in the development of effective mitigation strategies. In the realm of autonomous systems, models trained on vast amounts of data can enhance the safety and efficiency of self-driving cars, leading to reduced accidents and congestion. Large language models, like GPT-3, have shown remarkable language generation capabilities, enabling advancements in areas such as machine translation, content generation, and virtual assistants.

However, these benefits come with certain visible or hidden costs that we need to pay to reach them. Efforts are being made to address the environmental impact of AI. These include measuring and standardising carbon emissions, using renewable energy sources for data centres, and improving energy efficiency in computing systems. Awareness of the energy and environmental costs of AI is essential to make informed decisions about its usage. There are also significant social challenges around the use of AI systems: accessibility to these systems, legal concerns about training data and biases and impacts on the workforce. We need to address these challenges, too.

2.1 Environmental costs

The training process for these models requires vast computational resources, leading to substantial energy consumption. For instance, a study (Strubell et al., 2019) conducted by researchers at the University of Massachusetts, Amherst, found that training a state-of-the-art language model, similar to GPT-3, can emit approximately 626,000 pounds (284 t) of carbon dioxide (CO_2), equivalent to the emissions produced by five average passenger vehicles during their entire operational lifespan. Training a single BERT base model on GPUs requires about 1,500 kWh of energy, equivalent to a trans-American flight between New York and San Francisco. Another study (Amodei and Hernandez, 2018) showed that the amount of computing used to train the largest AI models has increased by 300,000x in 6 y. This leads to staggering energy consumption of AI systems despite the progress towards more efficient hardware solutions and better-optimised algorithms.

The cooling of data centres, which house the computational infrastructure required for training these models, adds to the environmental strain. Data centres consume significant amounts of water for cooling purposes. Water is primarily used in data centres for cooling systems to control the heat generated by the facilities and ensure uninterrupted operation. According to an article (Zhang, 2022), the amount of water consumed by data centres depends on factors such as facility size, cooling system, and outdoor temperature and humidity. A report (Koorney, 2021) by the US National Resources Defense Council (NRDC) states that US data centres alone consumed an estimated 626×10^9 L of water in 2020, equivalent to the annual water usage of over 2 million households. The high water usage in data centres, particularly from drinking water sources, raises concerns about its impact on local populations and the availability of water resources, especially in water-stressed regions. Bender et al. (2021) claim that the risks and benefits of this technology are unevenly distributed between communities: the negative effects of climate change are impacting the world's most marginalised communities the worst, and they raise the question of whether it is "fair or just to ask, for example, that the residents of the Maldives (likely to be underwater by 2100 or the 800,000 people in Sudan affected by drastic floods pay the environmental price of training and deploying ever larger English LMs, when similar large-scale models aren't being produced for Dhivehi or Sudanese Arabic?"

2.2 Strategies toward environmental sustainability

The environmental impact of AI model training has become a pressing concern, prompting researchers and industry leaders to explore strategies for reducing energy consumption and carbon emissions. Researchers have been working on developing more efficient algorithms that can achieve comparable performance with fewer computational resources. Additionally, specialised hardware such as tensor processing units (TPUs) has been designed specifically for accelerating AI computations, offering higher performance while consuming less power. Google, for instance, has reported that TPUs can provide up to 15 times higher performance per watt compared to traditional CPUs (Google AI Blog, 2018).

Another strategy focuses on using renewable energy sources to power data centres, minimising the carbon footprint associated with training AI models. Several tech giants and cloud service providers have made significant commitments to transition their data centres to renewable energy sources. As of 2021, Google reported that it matched its global energy consumption with 100 % renewable energy for its data centres and offices (Google Sustainability, 2021). Similarly, Microsoft announced plans to be carbon-negative by 2030 and rely on 100 % renewable energy for its data centres by 2025 (Microsoft, 2021).

Water cooling in data centres is highlighted as an effective method to reduce carbon emissions and increase sustainability (Hidalgo, 2022). Hyperscale data centre operators like Amazon Web Services (AWS), Microsoft, Google, and Facebook have committed to becoming water-positive by 2030, replenishing more water than they consume. AWS, for example, aims to improve water efficiency, use sustainable water sources, return water for

community reuse, and support water replenishment projects. Microsoft focuses on reducing water use intensity and replenishing water in water-stressed regions, while Google aims to replenish 120 % of the water it consumes by investing in community projects and improving watershed health.
These efforts are essential in mitigating the environmental impact of AI model training. By optimising algorithms, utilising specialised hardware, and transitioning to renewable energy sources, the energy consumption and carbon emissions associated with training large AI models can be significantly reduced.

2.3 Social challenges

The accessibility of these models is a critical factor in determining their social implications. Inadequate access can perpetuate the digital divide, limiting the opportunities for marginalised communities, low-income individuals, and those without advanced technical skills. The restricted access may hinder innovation and impede the development of solutions that address diverse societal challenges. The increasing capabilities of AI models can have implications for the labour market (Solaiman et al., 2020). As AI models become more sophisticated and capable, certain jobs may become automated or rendered obsolete. This can lead to job displacement and require efforts to reskill or retrain individuals for new roles in the job market.
In addition to accessibility, there are concerns regarding data privacy, biases, and ethical implications associated with the use of these models. AI models are trained on vast amounts of data, which can include personal information, leading to privacy concerns. Biases present in the training data can be propagated and amplified by these models, resulting in biased outputs and decisions. This can perpetuate existing societal biases and discrimination.

2.4 Possible solutions

Ensuring equitable access to large language models is essential to promote inclusivity, diversity, and fair opportunities for all, fostering a more just and participatory society. What can we do about this? Andrew Ng believes that “our best bet is to work quickly to democratise access to AI by (i) reducing the cost of tools and (ii) training as many people as possible to understand them. This will increase the odds that people have the skills they need to keep creating value. It will also ensure that citizens understand AI well enough to steer their societies toward a future that’s good for everyone” (Ng, 2023a).
We also believe that training in AI and data literacy for many more people is the way forward. Better training will enable people to solve a wider variety of problems, enriching society. Addressing these social and ethical challenges requires multidisciplinary collaboration among researchers, policymakers, industry stakeholders, and society. By considering ethical frameworks, implementing safeguards for data privacy and bias mitigation, and promoting inclusivity and responsible deployment, we can strive to harness the benefits of AI while mitigating its potential negative consequences.
Van Wynsberghe (2021) explores the concept of Sustainable AI in her paper titled “Sustainable AI: AI for sustainability and the sustainability of AI”. Sustainable AI is defined as a movement that aims to promote change throughout the entire lifecycle of AI products, including idea generation, training, implementation, and governance, with a focus on ecological integrity and social justice. Sustainable AI should encompass the principles of sustainable development, which involve inter- and intra-generational equity and the three interconnected pillars of the environment, economy, and society.

3. Bringing a sustainable future with regulations?

The need for regulations for managing AI has gained significant attention globally, and many nations have taken significant steps to address AI regulation. More closely, in the European Union (EU), lawmakers have been diligently working on the AI Act since 2021, aimed at regulating automated systems based on their potential for harm. This legislation is expected to become law soon. Additionally, the EU has created the European Centre for Algorithmic Transparency (ECAT), a regulatory body dedicated to studying algorithms used on social media sites and search engines. ECAT's role is to ensure compliance with the European Union's Digital Services Act, which aims to block online hate speech, targeted ads, and other objectionable content. The reports and audits submitted by companies to European regulators will be analysed by ECAT to assess algorithmic compliance.
The rapid development of new AI products has become possible with the aid of new tools and models, enabling quick and cost-effective iterations with users and without lengthy testing. However, this speed should not overshadow the importance of responsible AI development. Andrew Ng emphasised that developers must carefully evaluate the potential risks, such as bias, unfairness, privacy violations, or malicious use, before widely deploying AI systems to ensure their safety and avoid harmful consequences (Ng, 2023b). As AI capabilities continue to grow, concerns about its potential negative impacts have risen. Implementing thoughtful regulations and enforcement mechanisms can align AI development and application with societal benefits. For businesses, well-defined guidelines will be crucial in preventing harm to the public and protecting their reputation. Sam

Altman suggested that startups could be regulated more lightly than established companies due to their smaller reach and lower risk of harm. With time, these startups might grow, expand, and become subject to more stringent regulations, ensuring responsible and ethical AI practices.
In the future, AI systems are expected to be integrated into a wide range of products and services as advancements in algorithms continue to bring new solutions to the market. Consequently, the volume of data used for training and inference in AI models is projected to increase at an even faster pace. We foresee that an increasing portion of data generated by Internet of Things (IoT) devices is likely to be consumed by AI systems, contributing to the exponential growth in compute requirements, as illustrated by Amodei and Hernandez (2018), with a staggering 300,000x increase in just 6 y. Managing such unprecedented growth in AI infrastructure becomes crucial, and proactive measures must be taken now to mitigate its impact.

4. Regulatory framework: a proposal

We believe that regulations for AI are essential for several reasons, but currently, few regulators fully comprehend the potential benefits and risks of AI to create effective laws. Governments need a deeper understanding of technology before crafting regulations, and the establishment of ECAT is a positive step in that direction. The regulations should be flexible, promote innovation, and involve collaboration between the market economy and governments.
To assess the direct and indirect impact of AI services, we require greater visibility into large AI companies. In many countries, publicly traded companies must disclose significant financial information, while even smaller companies provide basic balance sheets and profit and loss statements. Companies might find these requirements intrusive, but the resulting transparency fosters trust. We can also draw parallels with good manufacturing practice guidelines that ensure the safety of manufactured products for human consumption or use. Similar standards can be developed for AI products and services, and countries should mandate large AI companies to follow generally accepted standards and disclose data about their activities in detail to guarantee adherence to the rules.
We propose setting up a green standard and good and ethical practices standard, possibly with different levels. Legislation could require that AI products meet specific levels under certain conditions, such as when they serve a large user base or operate in sensitive domains. The professional community and legislators could collaboratively create an innovation-friendly framework, allowing AI companies to incrementally meet higher standards as they grow or move into more sensitive areas.
Recent news (Nagle, 2023) indicates progress in this direction, at least in the US. Seven major AI companies pledged to allow external security testing of their products and share AI risk management data with governments, civil society, and academia. President Joe Biden commended these voluntary measures to enhance safety and transparency around emerging AI technology. However, he also noted that more steps are needed to ensure the safety of this evolving technology and called for new legislation.

4.1 Green data centres and certificates

As the global data centre market is projected to grow at 7.5 % annually in the next decade (Globe Newswire, 2021), it is essential to have greener data centres in the future. A green data centre is defined as a facility designed to be highly energy efficient and minimise its environmental impact while hosting servers for data storage, management, and dissemination.
To objectively measure the energy efficiency of a data centre, the following metrics have been developed:

- Power Usage Effectiveness (PUE): This metric, developed by The Green Grid, measures the ratio of power consumed by a data centre to the power delivered to its equipment. A sustainable data centre has a PUE ratio of 1.0, indicating that all power is efficiently utilised by the IT equipment.
- Carbon Usage Effectiveness (CUE): Also developed by The Green Grid, CUE evaluates a data centre's ratio of total CO_2 emissions to the energy consumption of IT equipment. A green data centre aims to achieve the lowest possible CUE value, indicating minimal carbon emissions.
- Water usage effectiveness (WUE) is a metric used to measure data centre water efficiency, calculated by dividing the annual water usage by the energy consumption of IT equipment. The industry average WUE in 2021 was 1.80 L/kWh of water per electricity used.

To establish a green data centre, operators should focus on two key strategies. Firstly, they should transition to renewable energy sources, and secondly, optimise energy efficiency to minimise carbon emissions. This involves carefully selecting cooling technology tailored to the specific needs and location of the data centre. Data centre operators are increasingly adopting renewable energy sources and entering into Power Purchase Agreements (PPAs) to procure renewable energy. Additionally, data centres can enhance their water efficiency by using non-potable water sources like greywater or recycled water. Water can be conserved and reused through recirculation in cooling systems, further promoting sustainability (Araner, 2023).

To encourage environmentally friendly practices, AI companies should conduct audits of their data centres, or alternatively. Data centres could acquire green certificates and provide them to the companies operating within their facilities. These certificates would describe the data centre's efficiency, measured using metrics like PUE, CUE or WUE, and other factors. The certificates could be categorised into different levels, indicating the degree of "environment friendliness." A standard with five levels could be established, and legislation could require large AI and internet companies, such as Google, Microsoft, and Amazon, serving billions of people, to meet the strictest standard. This approach would incentivise companies to prioritise sustainable practices and contribute to a greener future.

4.2 Good and ethical AI practices

To ensure the sustainability of AI from both social and economic perspectives, it is essential that we establish standards with comprehensive guidelines for building, testing, and monitoring AI systems, similar to good manufacturing and laboratory practices. The complexity of the domain requires expertise from various stakeholders, and regulatory oversight is also necessary. Industry and academia can contribute to formulating guidelines, and government regulations can reinforce widely accepted norms in the field.
In this regard, Google AI (2023) has already taken significant steps in shaping "Responsible AI practices," which encompass general recommended practices and specific measures for fairness, interpretability, privacy, and safety. Designing AI systems requires consideration of unique aspects related to machine learning. Google AI recommends using multiple metrics to assess training and monitoring, examining raw data, understanding dataset and model limitations, and maintaining monitoring and updates after deployment.

- Regarding fairness, Google AI emphasises the need to establish fairness and inclusion goals, especially if AI is used for more critical tasks like diagnosing medical conditions. Representative datasets, monitoring for unfair biases, and performance evaluation are essential aspects of this.
- Interpretability is crucial for questioning, understanding, and trusting AI systems. It enables scientists and engineers to design, develop, and debug models effectively. Ensuring interpretability should be an integral part of the user experience, with engineers having a good understanding of the trained model and the end goals and effectively communicating explanations to users.
- Privacy is a vital consideration, not only to adhere to legal and regulatory requirements but also to respect social norms and individual expectations. Google AI recommends responsible data collection and handling, leveraging on-device processing where applicable, and safeguarding the privacy of machine learning models.
- Safety is of utmost importance, particularly in safety-critical applications. Addressing safety challenges unique to AI systems, such as predicting unforeseen scenarios when ML is applied to problems that are difficult for humans to solve, is very difficult but more than necessary, especially in the era of generative AI.

While these are fundamental standards, more specific and detailed guidelines should be developed for various domains, such as social media, self-driving systems, medical diagnostics, and law enforcement. Establishing special audit organisations, like ECAT, should play a significant role in ensuring compliance and transparency. Large companies with substantial user bases or those operating in sensitive areas would be required to undergo audits and disclose detailed information about their systems and algorithms. A tiered approach with different levels of standards should be employed. Legislation could mandate adherence to specific standards above a certain user base or for applications in critical areas. An AI company with a certain user base (e.g., monthly active users above 100 million for a year) or serving special areas (e.g., face recognition for law enforcement, credit scoring for banks, diagnosing medical images in the health industry, etc.) would need to conform with the standard above level 4 for example.
To achieve this, disclosure details and data granularity need to be worked out. Shared data, analysed by independent organisations or government agencies, could provide insights into AI system behaviour. In the case of social media, the delivery of different-flavoured content to different user subsets could shed light on biases. For autonomous driving systems, the reliability of the solution under different conditions could be better assessed and facilitate better testing and refinement before market deployment. Specific standards could be formulated for other sensitive areas (like diagnosing illnesses based on medical images, etc.) Companies meeting these requirements would benefit by being able to use an AI label in their marketing communications or it could be a precondition to sell their products to major private or public organisations.

5. Conclusion

This article highlights the growing environmental and social concerns surrounding the use of large AI models. If used responsibly and ethically, AI models have the potential to generate enormous value and solve important problems for humankind. However, achieving this potential requires addressing challenges and ethical

considerations. The environmental costs associated with training and using these models are substantial (in 2018, data centres were estimated to consume about 1% of all global energy – and despite the success of energy efficiency measures, it is increasing). Strategies toward sustainability, such as optimising algorithms utilising renewable energy sources, exist and can help mitigate these impacts. Environmental impacts should be monitored, and regulations designed and enforced to tackle this problem. The paper proposes the wider adoption of green data centres and certificates to promote energy and water efficiency.

The social challenges related to AI include accessibility, data privacy, and fairness. The 300,000x times growth in the amount of computations clearly promises a widespread adoption of AI systems in many facets of our everyday lives. Ensuring equitable access to AI models and addressing these concerns require collaboration among researchers, policymakers, industry stakeholders, and society. To address these challenges, the article suggests the implementation of more comprehensive regulations and standards for AI development and usage. We call for guidelines on good and ethical AI practices formulated through joint efforts between government, industry, and academia. Such standards and regulations can promote responsible AI usage, inclusivity, transparency, and fairness while encouraging innovation and benefitting society.

References

Amodei D., Hernandez D., 2018, AI and Compute, OpenAI website. <openai.com/blog/ai-and-compute/>, accessed 22.10.2023.

Araner, 2023, What is a green data center?, Araner, <www.araner.com/blog/what-is-a-green-data-center>, accessed 22.10.2023.

Bender E.M., Gebru T., McMillan-Major A., Shmitchell S., 2021, On the dangers of stochastic parrots: Can language models be too big?, Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, March 3 - 10, 2021, Virtual Event Canada, 610-623.

Cho R., 2023, AI's Growing Carbon Footprint. State of the Planet, Columbia Climate School. <news.climate.columbia.edu/2023/06/09/ais-growing-carbon-footprint/>, accessed 22.10.2023.

Globe Newswire, 2021, Data Center Market – Growth, Trends, COVID-19 Impact, and Forecasts (2021 - 2026). Globe Newswire, <www.globenewswire.com/news-release/2021/02/05/2170186/0/en/Data-Center-Market-Growth-Trends-COVID-19-Impact-and-Forecasts-2021-2026.html>, accessed 22.10.2023.

Google AI Blog, 2018, The AIY Vision Kit, TensorFlow and TPUs: Everything you need to know, Google AI Blog <ai.googleblog.com/2018/05/the-aiy-vision-kit-tensorflow-and-tpus.html>, accessed 22.10.2023.

Google AI, 2023, Responsible AI practices, Google AI Blog <ai.google/responsibility/responsible-ai-practices/>, accessed 22.10.2023.

Google Sustainability, 2021, Our Progress, Google Sustainability <sustainability.google/projects/our-progress/>, accessed 22.10.2023.

Hidalgo M., 2022, Energy and water consumption in data centers: sustainability risks, IEEE Paper 69/2022 <www.ieee.es/en/Galerias/fichero/docs_analisis/2022/DIEEEA69_2022_MARHID_Datos_ENG.pdf>, accessed 22.10.2023.

Koorney J.G., 2021, How Much Water Do Data Centers Really Use? A new methodology and preliminary global results, <www.nrdc.org/sites/default/files/data-center-water-use-report-IP.pdf>, accessed 22.10.2023.

Microsoft, 2021, Microsoft will be carbon negative by 2030. Microsoft, <blogs.microsoft.com/blog/2020/01/16/microsoft-will-be-carbon-negative-by-2030/>, accessed 22.10.2023.

Nagle M., 2023, Biden White House, tech companies launch new safeguards around emerging AI technology. ABC News, <abcnews.go.com/Politics/biden-white-house-tech-companies-launch-new-safeguards/story?id=101555314>, accessed 22.10.2023.

Ng A., 2023a, AI Risk and the Resource Curse. The Batch, <www.deeplearning.ai/the-batch/ai-risk-and-the-resource-curse/>, accessed 22.10.2023.

Ng A., 2023b, Editorial letter for Issue 204. The Batch, <www.deeplearning.ai/the-batch/issue-204/>, accessed 22.10.2023.

Solaiman I., Brundage M., Clark J., Askell A., Herbert-Voss A., Wu J., Radford A., Krueger G., Kim J.W., Kreps S., McCain M., Newhouse A., Blazakis J., McGuffie K., Wang J., 2020, Release Strategies and the Social Impact of Language Models. <arxiv.org/abs/2004.04604>, accessed 22.10.2023.

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, Florence, Italy. July 2019, 3645–3650.

Van Wynsberghe A., 2021, Sustainable AI: AI for sustainability and the sustainability of AI. AI and Ethics, 1, 213-218.

Zhang M., 2022, Data Center Water Usage:Billions of Gallons Every Year. Dgtl Infra, December 8, 2022, <dgtlinfra.com/data-center-water-usage>, accessed 22/10/2023.