The carbon impact of artificial intelligence

The part that artificial intelligence plays in climate change has come under scrutiny, including from tech workers themselves who joined the global climate strike last year. Much can be done by developing tools to quantify the carbon cost of machine learning models and by switching to a sustainable artificial intelligence infrastructure.

Payal Dhar

In 2018, Kate Crawford and Vladan Joler's award-winning visual map and essay, titled 'Anatomy of an AI system', demonstrated the impact of an artificial intelligence (AI) device on a global scale in terms of human labour, data and resources that are required during its lifespan, from manufacture to disposal, using Amazon's Echo as an example. At every level, wrote the authors, contemporary technology is deeply rooted in the exploitation of human bodies. Starting from extracting metals from the Earth and the resulting environmental effects, to the sweatshops of programmers that keep the system going, to the personal data about the user that the device gathers, they offered a visual picture of AI's impact on the environment and human rights.

It has become an urgent matter to consider the role of AI technology in the climate crisis. The United Nations has called climate change a "defining crisis of our time", and, according to the Climate Reality Project, 97% of climate scientists concur that human activity is its main driver. The key mitigation pathways to avoiding a global environmental catastrophe include bringing emissions to zero by the middle of the twenty-first century, and limiting the average global warming to 1.5 °C. Prior to the COVID-19 crisis, this was deemed an eminently achievable goal should we act now; however, the pandemic is likely to have caused long-term implications that aren't yet clear.

AI seems destined to play a dual role. On the one hand, it can help reduce the effects of the climate crisis, such as in smart grid design, developing low-emission infrastructure, and modelling climate change predictions. On the other hand, AI is itself a significant emitter of carbon. This message reached the attention of a general audience in the latter half of 2019 when researchers at the University of Massachusetts Amherst analysed various natural language processing (NLP) training models available online to estimate the energy cost in kilowatts required to train them. Converting this energy consumption in approximate carbon emissions and electricity costs, the authors estimated that the carbon footprint of training a single big language model is equal to around 300,000 kg of carbon dioxide emissions. This is of the order of 125 round-trip flights between New York and Beijing, a quantification that laypersons can visualize.

Credit: Monty Rakusen/Cultura/gettyimages

But the carbon cost of training large machine learning models like the ones in the UMass study is only part of the problem; for a full picture, closer attention needs to be paid to the carbon impact of the infrastructure around big tech's deployment of AI. Last year saw tech workers urging their employers to recognize their part in the climate crisis. Thousands joined the global climate strike in September 2019 to raise attention to big tech's collaboration with fossil fuel companies and its part in the repression of climate refugees and frontline communities. Employees of Amazon, Google, Microsoft, Facebook, Twitter and others, organizing as the Tech Workers Coalition, marched to demand from their employers a promise to reduce emissions to zero by 2030, to not have contracts with fossil fuel companies, to stop funding climate change deniers, and to stop the exploitation of climate refugees and frontline communities.

Need to quantify

A main problem to tackle in reducing AI's climate impact is to quantify its energy consumption and carbon emission, and to make this information transparent. Crawford and Joler wrote in their essay that the material details of the costs of large-scale AI systems are vague in the social imagination, to the extent that a layperson might think that building a machine learning (ML)-based system is a simple task. Part of the enigma lies in the absence of a standard of measurement.

Alexandre Lacoste and colleagues worked on a study to make quantifying the carbon cost of ML easier for researchers. The emissions incurred in the training of a neural

Credit: Imaginima/E+/gettyimages

network model, they found, are related to the location of the training server and the energy grid it uses, the length of the training procedure, and the hardware on which the training takes place. They developed an emissions calculator to estimate the energy use, and the concomitant environmental impact, of training ML models. Alexandra Luccioni, a collaborator on that study, says that keeping an eye on the emissions level of AI is crucial for the near future. "This is definitely something that people are working [on], be it via more efficient GPUs [graphics processing units] or by buying renewable energy credits for the carbon that was produced by neural network training. Using renewable energy grids for training neural networks is the single biggest change that can be made. It can make emissions vary by a factor of 40, between a fully renewable grid and a fully coal grid." However, to make AI less polluting, she adds, it needs to become more of a mainstream conversation, including "getting researchers to divulge how much carbon dioxide was produced by their research, to reuse models instead of training them from scratch and by using more efficient GPUs."

The actionable recommendations from the UMass team are similar — for example, the authors encourage researchers to prioritize computationally efficient hardware and algorithms, to report training time and sensitivity to hyperparameters in published performance results, and to perform a cost–benefit analysis of NLP models for comparison.

Red AI

A 2018 analysis led by Dario Amodei and Danny Hernandez of the California-based OpenAI research lab, an organization that describes its mission as ensuring that artificial general intelligence benefits all of humanity, revealed that the compute used in various large AI training models had been doubling every 3.4 months since 2012 — a wild deviation from Moore's Law, which puts this at 18 months — accounting for a 300,000× increase. This directly corresponds to the advances seen in the AI industry in recent years. While algorithmic innovation and data — two other factors directly related to the growth of AI — are difficult to quantify, compute isn't. But, they wrote in their blog post, the nebulousness of the exact amounts of compute can and does just as easily function as a fig leaf to hide the shortcomings of current algorithms.

Roy Schwartz and collaborators, in a position paper published in mid-2019, called this trend 'red AI', that is, 'buying' stronger results by using massive compute. For a linear gain in performance, an exponentially larger model is required, which can come in the form of increasing the amount of training data or the number of experiments, thus escalating computational costs, and therefore carbon emissions. To demonstrate the prevalence of red AI, Schwartz et al. analysed over 60 papers from top conferences and concluded that a vast majority — 90% of papers from the 2018 Annual Meeting of the Association for Computational Linguistics, 80% from the 2018 Conference on Neurological Information Processing Systems (NeurIPS), and 75% from the 2019 Conference on Computer Vision and Pattern Recognition — prioritized accuracy over efficiency.

Their study enumerated three factors making AI research red: the cost of executing the model on a single example; the size of the training dataset, which controls the number of times the model is executed; and the number of hyperparameter experiments, which controls how many times the model is trained. The total cost of producing a result in ML, they said, increases linearly with each of these quantities. Overall, the increasing trend towards neural architecture search (NAS) and automated hyperparameter optimization (also called autoML), which are extremely heavy on compute, contribute substantially to red AI. In their paper, Schwartz and colleagues advocate for 'green AI', which they defined as "AI research that yields novel results without increasing computational cost, and ideally reducing it", opposite to red AI.

Despite all this, red AI can still be valuable. In pushing the limits of model size, dataset size and the hyperparameter search space, even if a massive amount of resources is required, this may still pay off in terms of downstream performance. Justin Burr, communications manager for Google AI, says, "Training better models can actually save more energy over time. For example, NAS can find very efficient models. Given the number of times they're used each day for inference, in less than a week they save more in energy than the old hand-tuned versions."

Into the grid

AI scientist Richard Sutton, often called the 'father of reinforcement learning', wrote a blog post in early 2019 titled 'The bitter lesson', saying that AI methods that leverage computation are better and more accurate than those that leverage human knowledge. This mindset divides the AI industry, but what is indisputable is that relying on increasing amounts of compute and data requires ever-increasing power and other infrastructure, which is directly proportional to a rising carbon footprint. For a full grasp of AI's carbon impact, it is not enough to scrutinize the compute costs incurred by training large models.

With tech companies reticent about sharing data, and having no incentives to do so, any attempt at quantifying emissions remains difficult. Roel Dobbe of the AI Now Institute at New York University says that this has led to a peculiar kind of complexity. The aggressive pace with which

the computer industry has globalized and consolidated to a few players has challenged many societies' ability to retain control over critical infrastructure.

"For the computing infrastructure to be effective and efficient in terms of being able to offer compute across the globe, data centres have to be built regionally. [But with the] market mostly dominated by three American players, they are building data centres not just in their own countries, but across the world. And this has various impacts." One of these, he goes on to say, is the need for the right checks and balances, including regulations, to retain some form of local agency over these infrastructures. "These companies also invest a lot of money in alarming lobbying against regulation and other mechanisms that balance against their own power." This also raises other questions. With so few players dominating the industry, how does that impact the price of compute? What are the strategies that other players might employ who do not necessarily care about or have to adhere to more stringent requirements for their energy mix?

Companies are hesitant to share data about their energy mix. Greenpeace's *Clicking Clean* report from 2017 says that many companies who had committed to a 100% renewable future are more in a state of status quo than on a transformational path — in fact, despite net-zero-by-2040 pledges, Amazon's emissions increased by 15% last year. The report also points out that despite a drive towards renewable energy in significant markets, in others there has been a concomitant push for fossil-fuel-based energy. One such example is Virginia, USA, the data centre hub of the world, where only 1% of electricity comes from renewable sources. Then, there is the nexus of Big Data with Big Oil, the report says. Amazon, Google, Microsoft, Royal Dutch Shell and many others market their AI solutions to companies that work on fossil fuel extraction and use.

Estimating the carbon footprint of AI technologies, says Dobbe, should be fairly straightforward. "It's hardware that's running and we know how many operations various algorithms need to run." He compares tech to the aerospace industry, where it is quite easy to track emissions: "We know the energy efficiency of planes because there are standards and there are reports about what hardware planes use, how far and how long they fly, et cetera… We need to get to a similar point for the computing industry, not just for data centres, but also for the rest of the network infrastructure. It is mostly a matter of political will and consumer awareness to enforce similar kinds of transparency."

"I think that more tax incentives should be given for cloud providers to open data centres in places with hydro or solar energy," says Alexandra Luccioni. "For example, in Quebec, we have a very low-carbon grid that relies mostly on hydro, plus with the cold winters, the heat generated by computing centres can be used to heat homes. If companies had a big incentive to build their data centres there and not in, say, Texas, where the grid is mostly coal-fuelled, it could make a massive impact."

Switch to green

The Copenhagen Centre on Energy Efficiency, a partnership between the United Nations Environment Programme and the Technical University of Denmark, is a research and advisory work on climate, energy and sustainable development. Gabriela Prata Dias, head of the centre, and Xiao Wang, programme officer, stress that environmental sustainability should be considered as one of the principles towards responsible development and application of AI: "It is important to note that AI is not only just a tool but a resource demander… [and] the benefits of using such technology should outweigh its drawbacks." They suggest various steps to apply cleaner AI practices. First, they say, the definition of green AI needs to be actionable for all the relevant stakeholders in the industry, rather than be an abstract concept to most technical experts. They advocate for the essential role of standards to drive green AI adoption. "Environmental standards should be developed to ensure the mitigation of environmental impacts…[and] green AI certifications could be introduced to facilitate the industry process for promoting green AI development. For the organizations and companies that are using and deploying AI technologies, practical industry framework and guidelines that support green procurement of AI technologies would support them in looking for environmentally friendly AI practices." Finally, they add, it is imperative for governments to consider the long-term impacts in setting up a regulatory frameworks and legislations in a way that would legally address transparency and sustainability in AI development.

Deepika Sandeep, an AI scientist who heads the AI and ML programme at Bharat Light & Power (BLP), a Bengaluru-based clean energy generation company, feels that judicious use of deep learning needs to be enforced. "Not every problem demands a machine learning-based solution… [Since] training is the one place which consumes a lot of computational power and hence [increases] carbon footprint, what we do [at BLP] is we minimize our training cycles. Once they are deployed in production there is no training to be done… retraining is done only once in three or six months, depending." Reaching for solutions based on deep neural networks and deep learning architectures to solve simple problems that can be solved by other, less compute-intensive AI, is "where we are messing with the environment".

Next level AI

Back in October 2019, Roel Dobbe and Meredith Whittaker, co-founder of the AI Now Institute, wrote a paper on AI and climate change where they advocated seven policy recommendations that could plot the first steps in "tech-aware climate policy, and climate-aware tech policy". These were: mandate transparency, account for the entire tech ecosystem, watch for rebound effects, make non-energy policy a standard practice, integrate tech and climate policy, curb the use of AI to extract fossil fuels, and address AI's impact on climate refugees. "In the end," says Dobbe, "a lot of what the climate fight is about is that we need solidarities, and people on the ground, across the world to making efforts to create this kind of transparency."

Perhaps Miguel Luengo-Oroz, AI strategist and chief data scientist at the United Nations Global Pulse, has the right idea — that it's time to get to the "next level", which is a bigger intersection of AI and climate science. Having attended both NeurIPS and the United Nations Climate Change Conference (COP25) in December 2019, he noticed that the intersection between the two conferences was almost non-existent. "I don't know if anyone else was involved in both… We [spoke] about AI and sustainability at COP25, together with other new emerging technologies. And then at NeurIPS, where there were a lot of researchers, they met few climate experts… We need real experts; experts who deeply understand both sides of the game." ❐

Payal Dhar
Freelance writer, New Delhi, India.
✉*e-mail: payal_dhar@yahoo.com*

Published online: 12 August 2020
https://doi.org/10.1038/s42256-020-0219-9