Rethinking Concerns About AI's Energy Use

By Daniel Castro | January 29, 2024

Concerns about the energy used by digital technologies are not new. Near the peak of the dot-com boom in the 1990s, a *Forbes* article lamented, "Somewhere in America, a lump of coal is burned every time a book is ordered online."[1] The authors of the article, which became widely cited in subsequent years in debates about energy policy, estimated that "half of the electric grid will be powering the digital-Internet economy within the next decade."[2] However, the estimate was wrong, with errors in both its facts and methodology.[3] In hindsight, there is no longer any dispute, as the International Energy Agency (IEA) estimates that today's data centers and data transmission networks "each account for about 1–1.5% of global electricity use."[4]

This mistake was not an isolated event. Numerous headlines have appeared over the years predicting that the digital economy's energy footprint will balloon out of control.[5] For example, as the streaming wars kicked off in 2019—with Apple, Disney, HBO, and others announcing video streaming subscription services to compete with Netflix, Amazon, and YouTube—multiple media outlets repeated claims from a French think tank that "the emissions generated by watching 30 minutes of Netflix is the same as driving almost 4 miles."[6] But again, the estimate was completely wrong (it is more like driving between 10 and 100 yards), resulting from a mix of flawed assumptions and conversion errors, which the think tank eventually corrected a year later.[7]

With the recent surge in interest in artificial intelligence (AI), people are once again raising questions about the energy use of an emerging technology. In this case, critics speculate that the rapid adoption of AI

combined with an increase in the size of deep learning models will lead to a massive increase in energy use with a potentially devastating environmental impact.[8] However, as with past technologies, many of the early claims about the consumption of energy by AI have proven to be inflated and misleading. This report provides an overview of the debate, including some of the early missteps and how they have already shaped the policy conversation, and sets the record straight about AI's energy footprint and how it will likely evolve in the coming years. It recommends that policymakers address concerns about AI's energy consumption by taking the following steps:

- Develop energy transparency standards for AI models.
- Seek voluntary commitments on energy transparency for foundation models.
- Consider the unintended consequences of AI regulations on energy use.
- Use AI to decarbonize government operations.

THE FACTS ABOUT AI'S ENERGY USAGE AND CARBON EMISSIONS

Creating accurate estimates of the energy use and carbon emissions of AI systems over their lifetimes is challenging because these calculations depend on many complex factors, including details about the chips, cooling systems, data center design, software, workload, and energy sources used for electricity generation. This problem is not unique to AI. As a group of energy researchers described the problem in an article in the *Annual Review of Energy and the Environment*:

> Creating credible estimates of electricity requirements for information technology is fraught with difficulty. The underlying data are not known with precision, the empirical data are limited, the most useful data are often proprietary, and the technology is changing so rapidly that even accurate data are quickly obsolete.[9]

However, several studies have attempted to quantify the current and future energy demands and carbon emissions of AI systems. Unfortunately, some of the initial estimates have fallen into the same trap as past early studies about the energy use of digital technologies and have produced misleading estimates. These studies generally consider the energy needed for an AI system over its lifetime in two stages: 1) training the AI model; and 2) using the AI model to respond to specific queries—a process called "inference."

Training AI Models

Researchers at the University of Massachusetts Amherst estimated in 2019 the carbon emissions of several AI models, one of the first major studies of its kind.[10] The study found that BERT—which at the time was Google's state-of-the-art large language model (LLM)—emitted

approximately 1,438 pounds of carbon dioxide (CO_2) during 79 hours of training using 64 advanced graphics processing units (GPUs), the chips commonly used for training AI models because of their superior parallel processing capabilities. To put this in perspective, a roundtrip flight from New York to San Francisco creates approximately 2,000 pounds of CO_2 emissions per passenger. The researchers also estimated carbon emissions for training an AI model for neural architecture search (NAS), a technique for automatically finding one or more neural network architectures for a given task—one of the most computationally complex problems in machine learning. Specifically, they evaluated the energy usage of a NAS used to create a better English-German machine translation model.[11] The researchers estimated that training the model in question generated 626,155 pounds of CO_2 emissions (roughly equivalent to 300 roundtrip flights from the East Coast to the West Coast).[12]

Not surprisingly, given journalistic tendencies to skew toward negative coverage of tech, virtually all the headlines in the popular media focused on this latter estimate despite its narrow use case.[13] Even respected scientific news outlets such as *MIT Technology Review* ran such headlines as "Training a single AI model can emit as much carbon as five cars in their lifetimes."[14] These articles suggested that the massive energy needed to train this particular AI model was normal despite this estimate clearly referring to an atypical example. It would be like an automotive news outlet running an article that suggested "driving a car emits as much carbon as an airplane" based only on a study that looked at the environmental impact of a flying car prototype.

Moreover, both the original research paper and the subsequent news articles often noted that while the large AI model outperformed existing ones at language translation benchmarks, the improvements were only marginal. The implication was that AI researchers are making trivial performance improvements at the expense of non-trivial amounts of carbon emissions. Indeed, other AI researchers made this point explicit in a widely read paper "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?"[15] They argued that it is "environmental racism" for wealthy Western nations to deploy ever-larger AI models because these AI systems will have negative impacts on poor communities in the Global South. Specifically, they wrote:

> Is it 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 [language models], when similar large-scale models aren't being produced for Dhivehi or Sudanese Arabic?[16]

Given the charges—that training AI systems is not only dangerous for the environment but also an overt act of racism—it is not surprising that many policymakers have raised questions about AI's energy consumption. However, the headline-making estimate in the 2019 study was wildly

incorrect—just like many prior claims about the oversized energy footprint of digital technologies. The University of Massachusetts Amherst researchers had made several false assumptions that grossly inflated their estimates both for the total energy used and the carbon emissions. In response to the 2019 study, the researchers involved in the NAS model provided a detailed summary of the energy use and carbon emissions from their work, noting why the outside researchers' estimates were wrong. The actual emissions were 88 times smaller than the earlier study's estimate.[17] Unfortunately, the popular media paid little attention to correcting the record or noting the new findings, and so the initial impressions have lived on.

Researchers have published multiple studies in subsequent years estimating the energy needed to train many well-known AI models as well as their carbon emissions. As shown in table 1, while larger models generally require more energy usage than smaller ones do, the exact figures vary significantly across different AI models. For example, researchers estimate that training GPT-3—the 175 billion parameter AI model used in the popular ChatGPT application—created 552 tCO_2 emissions, but comparable AI models including OPT (a 175 billion parameter AI model created by Meta) and Gopher (a 280 billion parameter AI model created by Google) have significantly smaller carbon footprints. Moreover, the efficiency of training AI models continues to improve. For example, 18 months after GPT-3, Google produced GLaM, an LLM with 1.2 trillion parameters. Despite GLaM being nearly 7 times larger than GPT-3 and outperforming the other AI model, GLaM required 2.8 times less energy to train.[18] Finally, the energy mix used to power the data center where developers train an AI model impacts its carbon emissions. For example, the developers of BLOOM used a French data center powered by nuclear energy, which reduced its carbon footprint.[19]

Despite the new research, groups critical of AI have repeatedly cited the initial incorrect study in their demands for policymakers to reduce investment in large-scale computing resources. For example, the American Civil Liberties Union (ACLU) sent a letter to the Office of Science and Technology Policy (OSTP) in October 2021 complaining about the "environmental costs" of the White House's planned National AI Research Resource (NAIRR) and arguing that "the NAIRR should focus on offering an alternative to the data- and compute-hungry applications that are the focus of many industry and research labs."[20] Similarly, the Center for AI and Digital Policy falsely claimed in 2022 that "AI-enabled systems require exponentially rising computing power. This increase in computing power requires substantial energy consumption, generating a huge carbon footprint and upending the green effects of digitalization."[21] In each case, they made these claims despite overwhelming evidence showing they were misleading and overblown.

Table 1: Estimated energy demand of training various AI models

Model	# of Parameters	Chips (model x #)	Hours	Energy (MWh)	CO_2 Emissions (metric tons)	Estimate Source
BERT	0.1B	V100x64	79	1.5	0.7	Strubell et al., 2019[22]
GPT-2	1.5B	TPUv3x32	168	1.7*	0.7*	Strubell et al., 2019[23]
Llama 2	7B	A100x(n/a)	N/A	74*	31.2	Meta, 2023[24]
Llama 2	13B	A100x(n/a)	N/A	147*	62.4	Meta, 2023
Llama 2	70B	A100x(n/a)	N/A	688*	291.4	Meta, 2023
LaMDA	137B	TPUv3x1024	1,385	451	26	Thoppilan et al., 2022[25]
GPT-3	175B	V100x10000	355	1,287	552	Patterson et al., 2021[26]
OPT	175B	A100x992	N/A	N/A	75	Zhang et al., 2022[27]
BLOOM	176B	A100x384	2,820*	433	24.7	Luccioni et al., 2022[28]
Gopher	280B	TPUv3x4096	920	1,151*	380	Rae et al, 2022[29]
PaLM	540B	TPUv4x6144 TPUv4x3072	1200 326	3,436*	271.4	Chowdery et al., 2022[30]
GLaM	1,162B	TPUv4x(n/a)	N/A	456	40	Patterson et al., 2022[31]
GPT-4	1,800B	A100x25000	2280	N/A	N/A	Walker, 2023[32]

* Inferred based on available data, see appendix for details

Using AI Models

Despite the attention from policymakers and the media on the energy costs of training AI models, multiple studies have concluded that most of the

energy costs associated with AI systems come from using AI models—a process known as "inference" (because the model is inferring results based on a given input). For example, Amazon Web Services estimates that 90 percent of the cost of an AI model comes from inference.[33] Similarly, a study from Schneider Electric estimates that 80 percent of the AI workload in data centers in 2023 is from inference and 20 percent is for training.[34] Finally, a study by researchers at Meta notes that the exact breakdown between training versus inference varies across use cases. For LLMs, they estimate that inference is associated with 65 percent of the carbon footprint, but for recommendation models where parameters must be updated frequently based on new data, they estimate an even split between training and inference.[35]

Multiple factors impact the amount of energy used during inference, including the type of task and the AI model. As shown in table 2, the energy requirements for inference can vary significantly by task. For example, using an AI model to classify text is generally computationally less intensive (and thus uses less energy) than using AI to generate an image.[36] Different AI models also have different energy costs, and within specific models (e.g., Llama 2 7B versus Llama 2 70B), a larger number of parameters generally requires more energy for inference.

Table 2: Average energy use per 1,000 queries by task[37]

Task	kWh
Text classification	0.002
Image classification	0.007
Object detection	0.038
Text generation	0.047
Summarization	0.049
Image generation	2.907

Given that training a particular AI model incurs a one-time cost, whereas using an AI model continues to consume energy over time, it makes sense that most of the energy used for AI will eventually come from inference. It also means that the energy requirements for running AI models will have a significant impact on the overall energy use for AI systems. While most critics have focused on the energy used to train AI models, some people have expressed concern about the energy used during inference.[38] For example, writing in the October 2023 edition of the journal *Joule*, one researcher estimated that interacting with an LLM requires approximately

10 times as much energy as conducting a typical web search query, and extrapolated from that estimate to conclude that "the worst-case scenario suggests Google's AI alone could consume as much electricity as a country such as Ireland (29.3 TWh per year)."[39]

There are many reasons to doubt that such a "worst-case" scenario is on the near horizon. In 2022, Google's total global energy consumption across the entire company was 21.8 TWh. [40] For the worst-case prediction to be true, Google's energy use for AI alone would have to more than exceed its current total global energy use. It is true that the company's energy consumption has grown over time, particularly from its data centers, as its business has grown. For example, Google's data centers used about 3 TWh more electricity in 2022 than the year before.[41] But while its overall energy usage has grown, for the three years between 2019 and 2021, the proportion of energy it used for machine learning remained constant—between 10 to 15 percent of its total energy consumption—with approximately 60 percent of that used for inference.[42]

One explanation for the relatively constant proportion of energy used for inference is the improvements seen in AI models and hardware. Indeed, as shown in table 3, both performance and efficiency tend to improve over time. The table shows that over a few years, the accuracy of computer vision AI models improved significantly. In addition, the energy requirements for inference across these models generally decreased with the release of a newer chip. As noted in one recent study of the energy used for inference in AI models, "when a SOTA [state-of-the-art] model is released it usually has a huge number of FLOPs [floating point operations], and therefore consumes a large amount of energy, but in a couple of years there is a model with similar accuracy but with much lower number of FLOPs."[43] In other words, the newest AI models may not be particularly efficient by design because researchers are focusing on performance improvements, but over time, researchers will address efficiency.

Table 3: Energy consumption for inference of deep neural networks used for computer vision for two different GPUs[44]

	Year	Top-1 Accuracy (ImageNet)	P100, released June 2016 (Joules)	V100, released June 2017 (Joules)
AlexNet	2012	56.52	0.033	0.023
GoogLeNet	2014	69.77	0.077	0.055
Vgg16	2014	71.59	0.542	0.373
ResNet50	2015	75.30	0.179	0.132

WHAT AI ENERGY FORECASTS GET WRONG

One reason forecasts about future energy demands from AI are so high is they use inaccurate or misleading measurements, as described previously. Another reason is the forecasts ignore the practical economic and technical realities that come with widespread commercialization of AI.

The Energy Use of AI Is Limited by Economic Considerations

Many of the high-end estimates for AI energy use are impractical because of the costs involved. Buying more chips, building more data centers, and powering those data centers is expensive. For example, as even the author of the prediction that Google's AI alone might consume 29.3 TWh annually admitted, reaching this level would require a $100 billion investment in chips along with billions more in operating costs for the data center and electricity.[45] Even large tech companies would find it unsustainable to pay for such massive amounts of computing. Businesses are profit-seeking enterprises and computing costs money; therefore, they are not going to offer services for long that cost more to operate than they receive in revenue. Either the energy costs for using AI will come down or how companies deploy AI will be limited by cost factors.

The Rate of Performance Improvements in AI Will Decline Over Time

AI models have improved significantly in the past few years. For example, OpenAI's LLM model, GPT-4, released in March 2023, can pass many popular exams designed for humans, such as the SAT, GRE, LSAT, and AP tests for a variety of subjects.[46] These results are a substantial improvement over its earlier model released the prior year. While AI still cannot perform many tasks as well as humans, such as abstract reasoning, now that some AI models perform so highly on many benchmarks, there is substantially less opportunity for improvement in certain domains. As a result, many developers will likely focus more on optimizing their AI models rather than squeezing out ever-smaller improvements in accuracy because they will not receive a return on investment for building and operating larger models.

Future Innovations Will Improve AI's Energy Efficiency

The history of computing is one of continuous innovation, and these innovations extend to energy efficiency. For example, over the past decade, demands on global data centers have increased substantially even as the energy intensity of data centers has decreased by approximately 20 percent annually.[47] Between 2010 and 2018, there was a 550 percent increase in compute instances and a 2,400 percent increase in storage capacity in global data centers, but only a 6 percent increase in global data center energy use.[48] These energy efficiency gains came from improvements in hardware, virtualization, and data center design, and they are part of the reason that cloud computing has been able to scale.

Similar trends are already appearing in AI. As one recent paper notes, "Many studies report that the size of neural networks is growing exponentially. However, this does not necessarily imply that the cost is also growing exponentially, as more weights could be implemented with the same amount of energy, mostly due to hardware specialization but especially as the energy consumption per unit of compute is decreasing."[49] Improvements in hardware and software will likely keep the pace of energy growth from AI in check. Chipmakers continue to create more efficient GPUs for AI. For example, Nvidia's recent transition from one generation of GPUs to another resulted not only in significantly faster processing but also nearly doubled energy efficiency.[50] Likewise, researchers continue to experiment with techniques, such as pruning, quantization, and distillation, to create more compact AI models that are faster and more energy efficient with minimal loss of accuracy.[51] These types of advancements are one reason the proportion of energy Google uses for AI has remained constant in recent years despite machine learning growing to account for 70 to 80 percent of the compute used at the company.[52] Indeed, as one researcher succinctly put it, "[AI's] energy consumption is not skyrocketing, contrary to commonly expressed fears."[53]

AI's Energy Footprint Ignores Substitution Effects

Discussing the energy usage trends of AI systems can be misleading without considering the substitution effects of the technology. Many digital technologies help decarbonize the economy by substituting moving bits for moving atoms. For example, sending an email replaces mailing a letter, streaming a movie replaces renting a DVD, and participating in a video conference replaces traveling to an in-person meeting. AI will have a similar impact over time, both by further digitalizing many activities (such as by improving the quality of video calls) and by using AI to complete tasks more efficiently than using human labor.

One study in 2023 estimates the carbon footprint of using AI versus using a human for writing a page of text or creating an illustration. After considering the carbon emissions for different AI models (ChatGPT, BLOOM, Midjourney, and DALLE-2) and comparing workers in the United States and India, the researchers found "AI writing a page of text emits 130 to 1,500 times less CO2e than a human doing so" and "AI creating an image emits 310 to 2,900 times less."[54] Of course, since using AI does not eliminate humans—they still exist and eat, breathe, etc.—using AI does not eliminate these carbon emissions. But AI does eliminate the carbon emissions from the devices humans use for these tasks, such as laptop or desktop computers. As shown in table 4, these savings can be substantial; however, there are limits to generalizing these findings. For example, by making it easier to produce text and images, the volume of activity may increase. Nevertheless, these findings show how, holding all else equal, using AI to substitute for human labor can reduce carbon emissions in certain cases.

Table 4: Carbon footprint (grams CO_2e) for using a human (in the United States) versus AI (BLOOM/Midjourney) for certain tasks

	AI	Laptop	Desktop	Human
Writing a page of text	0.95	27	72	1,400
Creating an image	1.90	100	280	5,500

HOW AI'S ENERGY USE FITS INTO THE BIGGER PICTURE

The debate about AI's energy use is part of a larger debate about how to address global climate change. Within that context, there are important factors policymakers should keep in mind.

AI Will Play an Important Role in Addressing Climate Change

There are many opportunities to use AI to reduce carbon emissions, support clean energy technologies, and address climate change. These opportunities span multiple industries, including the transportation, agriculture, and energy sectors. For example, AI is crucial for integrating renewable energy sources such as wind and solar into the electric grid by using data points to forecast supply and demand. Likewise, utilities are using AI for predictive maintenance of energy assets, managing and controlling grids, and setting dynamic pricing—all critical elements for an efficient electric grid.[55]

AI can also help make sense of complex climate data from sensors and satellites, such as changing sea levels, surface temperatures, and rainfall, to create better forecasts and address risks of climate change. For example, AI can detect methane emissions from satellite data, allowing regulators to more effectively monitor industry.[56] Similarly, farmers can use AI for precision agriculture, reducing their use of fertilizer and water and their associated environmental costs.[57]

Already businesses, governments, and consumers are using AI to operate more efficiently. AI is a key part of creating smart cities that use AI to operate efficient buildings, roads, waterways, and more.[58] For example, in California, the government is using AI to monitor over a thousand cameras to detect and respond to wildfires quickly, reducing carbon emissions that come from these fires.[59] And AI can enable firms to optimize industrial processes, reduce waste, and use energy more efficiently, thereby reducing their carbon intensity.[60] For example, logistics providers use AI to optimize delivery routes, thereby reducing fuel consumption of their fleets.[61] And in the consumer space, tools such as the Nest smart thermometers saved customers 113 billion KWh between 2011 and 2022 and more efficient driving from Google Maps has reduced carbon emissions by 1.2 million metric tons.[62]

As AI matures, policymakers should continue to look to the technology as a key tool for addressing climate change.

There Is No Unique Market Failure for AI's Energy Use

Solving the global climate challenge will require transitioning to clean energy technologies that have a price and performance on par with dirty ones.[63] In the interim, any activity that uses energy has an environmental impact, and AI's use of energy is no different. However, there are no unique market failures associated with AI's use of energy that would lead to greater environmental impact than alternative uses would. A kilowatt-hour used for AI is no different than a kilowatt-hour used for watching television, microwaving popcorn, powering lights, or any other activity. Indeed, as noted previously, in many cases, AI applications will be used as a substitute for less energy-efficient activities and to address climate change. In both cases, those who consume energy must pay for it, and rational actors will generally seek to minimize these costs. While such costs may not include negative externalities associated with energy use, that problem is not unique to AI and cannot be addressed for AI alone.

Large Tech Companies Have Made Bold Net-Zero Commitments

Large tech companies are at the forefront of AI, and these are the same companies that have made some of the boldest commitments among corporations to reducing their carbon footprints. Consider the following:

- Google (now Alphabet) became carbon neutral in 2007, the first major company to do so.[64] A decade later, it became the first major company to purchase enough renewable energy to match its electricity consumption.[65] And, in 2020, it purchased carbon offsets to eliminate the company's entire carbon legacy.[66] The company continues to press forward and has committed to operating all its data centers and campuses on carbon-free energy by 2030.[67]

- Amazon co-founded The Climate Pledge in 2019, whereby companies commit to net-zero carbon emissions by 2040, 10 years ahead of the Paris Agreement.[68] And, in 2022, Amazon reported that 90 percent of the electricity it consumed came from renewable sources, and it was on track to reach 100 percent by 2025, five years ahead of its goal of 2030.[69]

- Microsoft has committed to be carbon negative by 2030 and eliminate its company's entire carbon legacy by 2050.[70] It has also committed to using 100 percent renewable energy by 2025.[71]

- Facebook (now Meta) reached zero carbon emissions in its direct operations (i.e., data centers and offices) in 2020 and has pledged to reach net-zero emissions across its entire value chain by 2030.[72]

These companies have remained publicly committed to these pledges even as they lead in the development and deployment of AI. Indeed, their net-zero pledges are one of the reasons these companies must carefully consider the efficiency of the AI models they train and deploy.

HOW POLICYMAKERS SHOULD ADDRESS AI'S ENERGY USE

Witnessing the rapid advancements in AI, policymakers around the world are considering whether and how they should regulate the technology, including its energy usage. For example, UNESCO's "Recommendation on the Ethics of AI"—which was adopted by 193 member states in November 2021—states, "Member States and business enterprises should assess the direct and indirect environmental impact throughout the AI system life cycle, including, but not limited to, its carbon footprint, energy consumption … and reduce the environmental impact of AI systems and data infrastructures."[73] The impact of AI on energy and the environment should be part of the policy debate, but policymakers should also be careful not to overreact, especially given the prevalence of misleading narratives falsely depicting AI's energy consumption as out of control.

There are reasonable steps policymakers can take to ensure AI is part of the solution, not part of the problem, when it comes to the environment. To that end, policymakers should do the following:

Develop Energy Transparency Standards for AI Models

It is usually easier to manage things that can be measured, and energy use of AI models is no different. While many AI developers have begun publishing model cards—short documents that accompany the release of an AI model that detail information about its performance, limitations, and other relevant information—these do not always contain information about the energy used to train or use them.[74] When they do include information about the environmental impact, they tend to focus on the carbon emissions from training rather than the energy needs for inference.[75] This focus on the energy used for training is partially out of necessity, as the developers of a model cannot control the hardware others might run their model on in the future or specific use cases. But the amount of energy used for training does not necessarily impact the energy that will be required for inference. Therefore, choosing models based on the amount of energy used to train them rather than the life cycle energy costs could lead to less-efficient outcomes.

To address this problem, policymakers should support the development of energy transparency standards for AI models, both for training and inference. In the United States, for example, the National Institute of Standards and Technology should work with the Department of Energy to develop a recommended best practice for assessing the training and inference energy costs. For example, this standard might include a set of benchmark tests and hardware to give comparable energy performance

metrics across different models. The United States should also work with the G7 and the Organization for Economic Cooperation and Development (OECD) to ensure broad adoption of these energy transparency standards to avoid different disclosure practices in different jurisdictions, especially since some countries might make such disclosures mandatory.

Seek Voluntary Commitments on Energy Transparency for Foundation Models

While developing transparency standards for AI models will help, it will also be important for leading AI companies to adopt these standards and disclose this information publicly. The White House has proactively sought out and obtained voluntary commitments from most of the leading U.S.-based AI companies to promote "safe, secure, and transparent development and use of generative AI (foundation) model technology."[76] While these commitments included important pledges from companies to engage in extensive testing to detect vulnerabilities and promises to avoid discrimination and bias in their models, they did not include any commitments around energy. To address that shortcoming, the White House should continue its dialogue with these companies to seek a voluntary commitment to publicly disclose the energy required to train and operate these foundation models, as well as the associated carbon emissions, especially for cloud-based AI service providers. Making this information publicly available will give users of foundation models the option to take the environmental footprint of AI into consideration when deciding which AI services to use.

Consider Unintended Consequences of AI Regulations on Energy Use

Many policymakers have called on developers to ensure their AI models minimize bias, avoid hate speech, limit disclosure of private information, and align to other, often worthwhile, goals. In many cases, developers are actively working to create models and build safeguards to address these concerns because they have strong market incentives to do so. However, policymakers rarely consider that their demands can raise the energy requirements to train and use AI models. For example, debiasing techniques for LLMs frequently add more energy costs in the training and fine-tuning stages.[77] Similarly, implementing safeguards to check that LLMs do not return harmful output, such as offensive speech, can result in additional computing costs during inference.[78] Thus, many of the proposed mandates for AI models could come at the expense of energy efficiency goals. The converse is also true: Mandates for energy-efficient AI models could create trade-offs that result in AI models that are less fair and more biased than they otherwise might be.

The point is not that policymakers should never regulate any AI system, but rather that they should avoid rushing to regulate until they fully understand the implications of their decisions. For example, the EU's AI Act initially included no requirements around energy efficiency. However, in response

to some of the misleading claims about AI's environmental impact, the European Parliament's proposed revisions to the legislation included substantial additions around energy, such as directing AI developers to integrate "state-of-the art methods and relevant applicable standards to reduce the energy use, resource use and waste, as well as to increase their energy efficiency and the overall efficiency of the system."[79] These requirements were in tension with other obligations in the AI Act to eliminate bias from AI models. While the AI Act now only includes more reasonable energy transparency requirements, the proposal from the European Parliament shows the potential for bad facts to lead to bad policy.

Use AI to Decarbonize Government Operations

AI offers important opportunities to improve the quality and efficiency of many government services, and adopting AI broadly across government agencies at every level should be a key priority for policymakers. In addition, AI can help the public sector reduce carbon emissions through more efficient digital services, smart cities and buildings, intelligent transportation systems, and other AI-enabled efficiencies. At the 2022 United Nations Climate Change Conference of the Parties (COP27), the United States launched the Net-Zero Government Initiative, which commits national governments to reaching net-zero carbon emissions for their operations by 2050.[80] To accelerate the use of AI across government agencies toward this goal, the president should sign an executive order directing the Technology Modernization Fund—a relatively new funding system for federal government IT projects—to include environmental impact as one of the core priority investment areas for projects to fund. In addition, the United States should invite and share best practices for using AI in government from the other countries that are part of the Net-Zero Government Initiative.

CONCLUSION

Policymakers need accurate information about the energy implications of AI. Unfortunately, groups that oppose AI, whether from honest misunderstanding of the evidence or intentional cherry-picking of the facts, continue to push the narrative that AI's energy footprint is growing out of control. In December 2023—more than two years after the record was corrected—a columnist writing in *The Guardian* repeated the original false and misleading statistic about AI's energy impact that has generated so much concern. The article stated:

> A study in 2019, for example, estimated the carbon footprint of training a single early large language model (LLM) such as GPT-2 at about 300,000kg of CO2 emissions—the equivalent of 125 round-trip flights between New York and Beijing. Since then, models have become exponentially bigger and their training footprints will therefore be proportionately larger.[81]

Just as the early predictions about the energy footprints of e-commerce and video streaming ultimately proved to be exaggerated, so too will those estimates about AI likely be wrong. But given the enormous opportunities to use AI to benefit the economy and society—including transitioning to a low-carbon future—it is imperative that policymakers and the media do a better job of vetting the claims they entertain about AI's environmental impact.

APPENDIX

Conversions and inferences shown for Table 1. All figures from source cited in table, unless otherwise noted.

BERT

0.65 metric tons CO_2e = 1,438 lbs

GPT-2

1.7 MWh = 32 GPUs x 168 hours x 289W* x 1.1 PUE*

**Using measurements from LaMDA. See Thoppilan et al., 2022.*

0.7 metric tons CO_2e = 1.7 MWh x 0.429 CO_2e/KWh*

** Using measurement of U.S. average data center net CO_2e/KWh from Patterson et al., 2021.*

Llama 2

74 MWh = 184,320 GPU hours x 400W

147 MWh = 368,640 GPU hours x 400W

688 MWh = 1,720,320 GPU hours x 400W

BLOOM

2,820 hours = 1,082,990 million GPU hours / 384 GPUs

Gopher

1,151 MWh = 4,096 GPUs x 920 hours x 283W x 1.08 PUE

PaLM

3,436 MWh = [(6,144 GPUs x 1,200 hours) + (3,072 GPUs x 336 hours)] x 378.5W x 1.08 PUE

REFERENCES

1. Peter W. Huber and Mark Mills, "Dig more coal — the PCs are coming," *Forbes*, May 31, 1999, https://www.forbes.com/forbes/1999/0531/6311070a.html.
2. Ibid.
3. Jonathan Koomey et al., "Sorry, Wrong Number: The Use and Misuse of Numerical Facts in Analysis and Media Reporting of Energy Issues," Annual Review of Energy and the Environment 27, no. 1 (November 1, 2002): 119–58, https://doi.org/10.1146/annurev.energy.27.122001.083458.
4. Vida Rozite, "Data Centres and Data Transmission Networks," IEA, July 11, 2023, https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks.
5. Colin Cunliff, "Beyond the Energy Techlash: The Real Climate Impacts of Information Technology," Information Technology and Innovation Foundation, July 6, 2020, https://itif.org/publications/2020/07/06/beyond-energy-techlash-real-climate-impacts-information-technology/.
6. George Kamiya, "Factcheck: What is the carbon footprint of streaming video on Netflix?" *Carbon Brief*, February 25, 2020, https://www.carbonbrief.org/factcheck-what-is-the-carbon-footprint-of-streaming-video-on-netflix/.
7. Ibid.
8. Alex De Vries, "The Growing Energy Footprint of Artificial Intelligence," *Joule* 7, no. 10 (October 1, 2023): 2191–94, https://doi.org/10.1016/j.joule.2023.09.004.
9. Koomey et al., "Sorry, Wrong Number: The Use and Misuse of Numerical Facts in Analysis and Media Reporting of Energy Issues."
10. Emma Strubell, Ananya Ganesh, and Andrew McCallum, "Energy and Policy Considerations for Deep Learning in NLP," arXiv.org, June 5, 2019, https://arxiv.org/abs/1906.02243.
11. David R. So, Chen Liang, and Quoc V. Le, "The Evolved Transformer," arXiv.org, January 30, 2019, https://arxiv.org/abs/1901.11117.
12. Strubell, Ganesh, and McCallum, "Energy and Policy Considerations for Deep Learning in NLP."
13. Doug Allen and Daniel Castro, "Why So Sad? A Look at the Change in Tone of Technology Reporting From 1986 to 2013" (Information Technology and Innovation Foundation, February 2017), https://itif.org/publications/2017/02/22/why-so-sad-look-change-tone-technology-reporting-1986-2013/.
14. Karen Hao, "Training a Single AI Model Can Emit as Much Carbon as Five Cars in Their Lifetimes," *MIT Technology Review*, December 7, 2020, https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes.
15. Emily M. Bender et al., "On the Dangers of Stochastic Parrots," *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, March 2021, 610–23, https://doi.org/10.1145/3442188.3445922.

16. Ibid.

17. David A. Patterson et al., “Carbon Emissions and Large Neural Network Training,” arXiv.org, April 21, 2021, https://doi.org/10.48550/arxiv.2104.10350.

18. David S. Patterson et al., “The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink,” *IEEE Computer* 55, no. 7 (July 1, 2022): 18–28, https://doi.org/10.1109/mc.2022.3148714.

19. Melissa Heikkilä, “We’re Getting a Better Idea of AI’s True Carbon Footprint,” *MIT Technology Review*, November 15, 2022, https://www.technologyreview.com/2022/11/14/1063192/were-getting-a-better-idea-of-ais-true-carbon-footprint/.

20. “RE: Request for Information on an Implementation Plan for a National Artificial Intelligence Research Resource,” American Civil Liberties Union, October 1, 2021, https://www.aclu.org/wp-content/uploads/document/ACLU_Response_to_NAIRR_RFI.pdf.

21. “Response to the RFI request on the National Artificial Intelligence Research and Development Strategic Plan,” Center for AI and Digital Policy, March 4, 2022, https://www.caidp.org/app/download/8378181763/CAIDP-Statement-OSTP-03042022.pdf.

22. Strubell, Ganesh, and McCallum, “Energy and Policy Considerations for Deep Learning in NLP.”

23. Ibid.

24. “Llama 2 Model Details,” Facebook Research, October 14, 2023, https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md.

25. Romal Thoppilan et al., “LAMDA: Language Models for Dialog Applications,” *arXiv.Org*, January 20, 2022, https://doi.org/10.48550/arxiv.2201.08239.

26. Patterson et al., “Carbon Emissions and Large Neural Network Training.”

27. Susan Zhang et al., “OPT: Open Pre-Trained Transformer Language Models,” *arXiv.Org*, May 2, 2022, https://doi.org/10.48550/arxiv.2205.01068.

28. Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat, “Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model,” *arXiv.Org*, November 3, 2022, https://doi.org/10.48550/arxiv.2211.02001.

29. Jack W. Rae et al., “Scaling Language Models: Methods, Analysis & Insights from Training Gopher,” *arXiv.Org*, December 8, 2021, https://doi.org/10.48550/arxiv.2112.11446.

30. Aakanksha Chowdhery et al., “PaLM: Scaling Language Modeling with Pathways,” *arXiv.Org*, April 5, 2022, https://doi.org/10.48550/arxiv.2204.02311.

31. Patterson et al., “The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink.”

32. Stephen M. Walker, “Everything We Know About GPT-4,” Klu, September 1, 2023, https://klu.ai/blog/gpt-4-llm.

33. Benoit de Chateauvieux et al., “Optimize AI/ML workloads for sustainability: Part 3, deployment and monitoring,” AWS Architecture Blog, April 27, 2022, https://aws.amazon.com/blogs/architecture/optimize-ai-ml-workloads-for-sustainability-part-3-deployment-and-monitoring/.

34. Victor Avelar et al., “The AI Disruption: Challenges and Guidance for Data Center Design,” Schneider Electric, White Paper 110, Version 2.1., n.d., accessed December 1, 2023, https://download.schneider-electric.com/files?p_Doc_Ref=SPD_WP110_EN.

35. Carole-Jean Wu et al., “Sustainable AI: Environmental Implications, Challenges and Opportunities,” *arXiv.Org*, October 30, 2021, https://doi.org/10.48550/arxiv.2111.00364.

36. Alexandra Sasha Luccioni, Yacine Jernite, and Emma Strubell, “Power Hungry Processing: Watts Driving the Cost of AI Deployment?,” *arXiv*, November 28, 2023, https://doi.org/10.48550/arxiv.2311.16863.

37. Ibid.

38. Ibid.

39. De Vries, “The Growing Energy Footprint of Artificial Intelligence.”

40. “Google Environmental Report: 2023,” Google, July 2023, https://www.gstatic.com/gumdrop/sustainability/google-2023-environmental-report.pdf.

41. Ibid.

42. Patterson et al., “The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink.”

43. Radosvet Desislavov, Fernando Martínez-Plumed, and José Hernández-Orallo, “Trends in AI Inference Energy Consumption: Beyond the Performance-vs-Parameter Laws of Deep Learning,” *Sustainable Computing: Informatics and Systems* 38 (April 1, 2023): 100857, https://doi.org/10.1016/j.suscom.2023.100857.

44. Ibid.

45. De Vries, “The Growing Energy Footprint of Artificial Intelligence.”

46. “GPT-4,” OpenAI, March 14, 2023, https://openai.com/research/gpt-4.

47. Eric Masanet et al., “Recalibrating Global Data Center Energy-Use Estimates,” *Science* 367, no. 6481 (February 28, 2020): 984–86, https://doi.org/10.1126/science.aba3758.

48. Ibid.

49. Desislavov, Radosvet, Fernando Martínez-Plumed, and José Hernández-Orallo. "Trends in AI Inference Energy Consumption: Beyond the Performance-vs-parameter Laws of Deep Learning." Sustainable Computing: Informatics and Systems 38, (2023): 100857, accessed September 22, 2023. https://doi.org/10.1016/j.suscom.2023.100857.

50. “Power Efficiency,” Nvidia, n.d., accessed December 15, 2023, https://www.nvidia.com/en-us/glossary/power-efficiency/.

51. Torsten Hoefler et al., “Sparsity in Deep Learning: Pruning and Growth for Efficient Inference and Training in Neural Networks,” *arXiv.Org*, January 31, 2021, https://doi.org/10.48550/arxiv.2102.00554.

52. Patterson et al., “The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink.”

53. Ibid.

54. Bill Tomlinson et al., “The Carbon Emissions of Writing and Illustrating Are Lower for AI than for Humans,” *arXiv (Cornell University)*, March 8, 2023, https://doi.org/10.48550/arxiv.2303.06219.

55. Vida Rozite, Jack Miller, and Sungjin Oh, “Why AI and energy are the new power couple,” IEA, November 2, 2023, https://www.iea.org/commentaries/why-ai-and-energy-are-the-new-power-couple.

56. Vít Růžička et al., “Semantic Segmentation of Methane Plumes with Hyperspectral Machine Learning Models,” *Scientific Reports* 13, no. 1 (November 17, 2023), https://doi.org/10.1038/s41598-023-44918-6.

57. David Rolnick et al., “Tackling Climate Change with Machine Learning,” *ACM Computing Surveys* 55, no. 2 (February 7, 2022): 1–96, https://doi.org/10.1145/3485128.

58. Lin Chen et al., “Artificial Intelligence-Based Solutions for Climate Change: A Review,” *Environmental Chemistry Letters* 21, no. 5 (June 13, 2023): 2525–57, https://doi.org/10.1007/s10311-023-01617-y.

59. Daniel Trotta, “California Turns to AI to Help Spot Wildfires,” *Reuters*, August 11, 2023, https://www.reuters.com/world/us/california-turns-ai-help-spot-wildfires-2023-08-11/.

60. Juan Li et al., “The Impact of Artificial Intelligence on Firms’ Energy and Resource Efficiency: Empirical Evidence from China,” *Resources Policy* 82 (May 1, 2023): 103507, https://doi.org/10.1016/j.resourpol.2023.103507.

61. Chen et al., “Artificial Intelligence-Based Solutions for Climate Change: A Review.”

62. “Google Environmental Report: 2023,” Google, July 2023, https://www.gstatic.com/gumdrop/sustainability/google-2023-environmental-report.pdf.

63. Robin Gaster, Robert D. Atkinson, and Ed Rightor, “Beyond Force: A Realist Pathway through the Green Transition,” ITIF, August 29, 2023, https://itif.org/publications/2023/07/10/beyond-force-a-realist-pathway-through-the-green-transition/.

64. Urs Hölzle, “Carbon neutrality by end of 2007,” Google, June 19, 2007, https://blog.google/outreach-initiatives/sustainability/carbon-neutrality-by-end-of-2007/.

65. Urs Hölzle, “Meeting our match: Buying 100 percent renewable energy,” Google, April 4, 2018, https://blog.google/outreach-initiatives/environment/meeting-our-match-buying-100-percent-renewable-energy/.

66. Sundar Pichai, “Our third decade of climate action: Realizing a carbon-free future,” Google, September 14, 2020, https://blog.google/outreach-initiatives/sustainability/our-third-decade-climate-action-realizing-carbon-free-future/.

67. “Net-zero carbon,” Google, n.d., https://sustainability.google/operating-sustainably/net-zero-carbon/ (accessed January 8, 2024).

68. “Driving Climate Solutions,” Amazon, n.d., accessed January 8, 2024, https://sustainability.aboutamazon.com/climate-solutions.

69. Ibid.

70. "The 2023 Microsoft Impact Summary," Microsoft, October 2023, https://www.microsoft.com/en-us/corporate-responsibility/impact-summary.

71. Brad Smith, "Microsoft will be carbon negative by 2030," Microsoft, January 16, 2020, https://blogs.microsoft.com/blog/2020/01/16/microsoft-will-be-carbon-negative-by-2030/.

72. "Climate," Facebook, n.d., accessed January 8, 2024, https://sustainability.fb.com/climate/.

73. "Recommendation on the Ethics of Artificial Intelligence," UNESCO, 2022, https://unesdoc.unesco.org/ark:/48223/pf0000381137.

74. Margaret Mitchell et al., "Model Cards for Model Reporting," *arXiv.Org*, January 29, 2019, https://doi.org/10.1145/3287560.3287596.

75. "About Model Cards," Hugging Face, n.d., accessed January 8, 2024, https://huggingface.co/spaces/huggingface/Model_Cards_Writing_Tool.

76. "Voluntary AI Commitments," White House, September 2023, https://www.whitehouse.gov/wp-content/uploads/2023/09/Voluntary-AI-Commitments-September-2023.pdf.

77. Marius Hessenthaler et al., "Bridging Fairness and Environmental Sustainability in Natural Language Processing," *arXiv.Org*, November 8, 2022, https://doi.org/10.48550/arxiv.2211.04256.

78. Traian Rebedea et al., "NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails," *arXiv.Org*, October 16, 2023, https://doi.org/10.48550/arxiv.2310.10501.

79. "AI Mandates," EU Artificial Intelligence Act, June 20, 2023, https://artificialintelligenceact.eu/wp-content/uploads/2023/08/AI-Mandates-20-June-2023.pdf.

80. "CEQ Launches Global Net-Zero Government Initiative, Announces 18 Countries Joining U.S. to Slash Emissions from Government Operations," White House, November 17, 2022, https://www.whitehouse.gov/ceq/news-updates/2022/11/17/ceq-launches-global-net-zero-government-initiative-announces-18-countries-joining-u-s-to-slash-emissions-from-government-operations/.

81. John Naughton, "Why AI Is a Disaster for the Climate," *The Observer*, December 23, 2023, https://www.theguardian.com/commentisfree/2023/dec/23/ai-chat-gpt-environmental-impact-energy-carbon-intensive-technology.

ABOUT THE AUTHOR

Daniel Castro is the director of the Center for Data Innovation and vice president of the Information Technology and Innovation Foundation. He has a B.S. in foreign service from Georgetown University and an M.S. in information security technology and management from Carnegie Mellon University.

ABOUT THE CENTER FOR DATA INNOVATION

The Center for Data Innovation studies the intersection of data, technology, and public policy. With staff in Washington, London, and Brussels, the Center formulates and promotes pragmatic public policies designed to maximize the benefits of data-driven innovation in the public and private sectors. It educates policymakers and the public about the opportunities and challenges associated with data, as well as technology trends such as open data, artificial intelligence, and the Internet of Things. The Center is part of the Information Technology and Innovation Foundation (ITIF), a nonprofit, nonpartisan think tank.

Contact: info@datainnovation.org

datainnovation.org