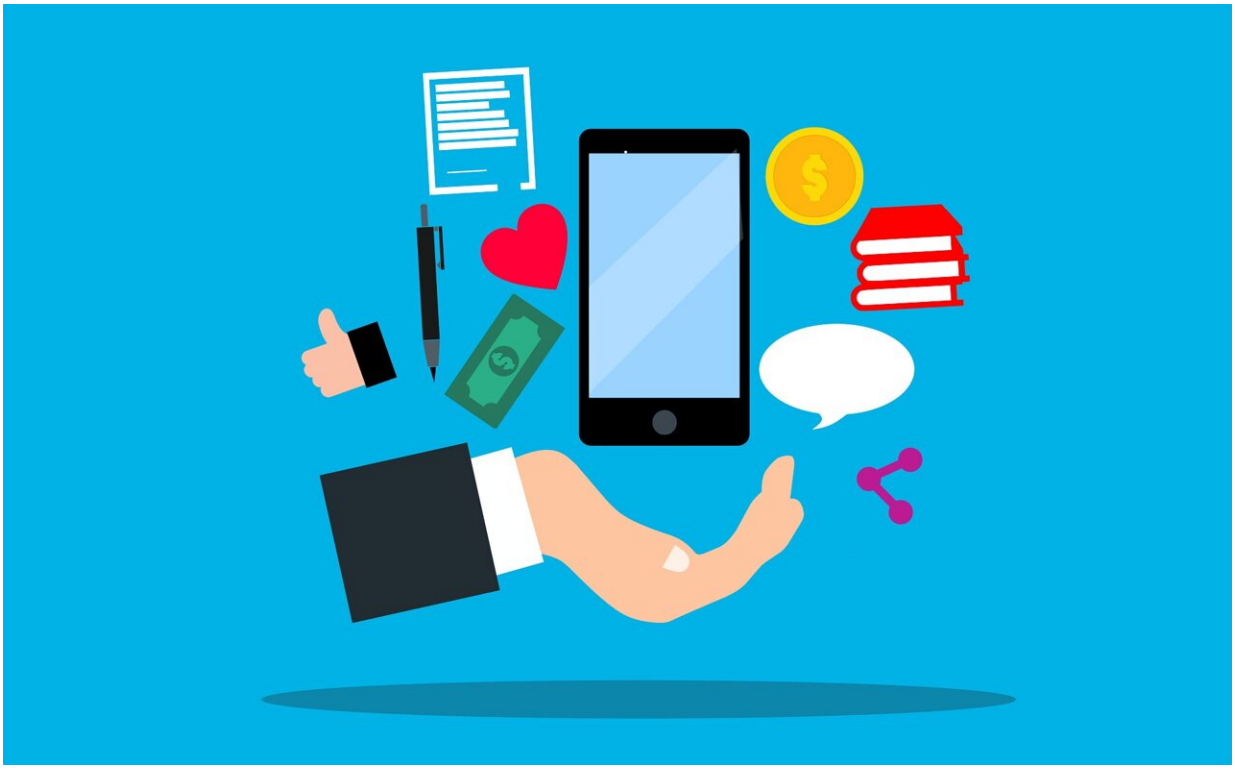
The hidden costs of AI: Impending energy and resource strain

March 9 2023, by Nathi Magubane



Credit: Pixabay/CC0 Public Domain

New technologies like the rapidly advancing deep learning models have led to increasingly sophisticated artificial intelligence (AI) models. With promises ranging from autonomous vehicles—land, air, and seafaring—to highly specialized information retrieval and creation like

ChatGPT, the possibilities seem boundless. Yet potential pitfalls exist, such as job displacement and privacy concerns, as well as materials and energy concerns.

Every operation a computer performs corresponds to electrical signals that travel through its hardware and consume power. The School of Engineering and Applied Science's Deep Jariwala, assistant professor of electrical and systems engineering, and Benjamin C. Lee, professor of electrical and systems engineering and computer and information science, spoke with Penn Today about the impact an increasing AI computation reliance will have as infrastructure develops to facilitate its ever-growing needs.

What sets AI and its current applications apart from other iterations of computing?

Jariwala: It's a totally new paradigm in terms of function. Think back to the very first computer, the Electrical Numerical Integrator and Computer (ENIAC) we have here at Penn. It was built to do math that would take too long for humans to calculate by hand and was mostly used for calculating ballistics trajectories, so it had an underlying logic that was straightforward: addition, subtraction, multiplication, and division of, say, 10-digit numbers that were manually input.

Lee: Computing for AI has three main pieces. One is data pre-processing, which means organizing a large dataset before you can do anything with it. This may involve labeling the data or cleaning it up, but basically you're just trying to create some structure in it.

Once preprocessed, you can start to "train" the AI; this is like teaching it how to interpret the data. Next, we can do what we call AI inference, which is running the model in response to user queries.

Jariwala: With AI, it's less about crunching raw numbers and more about using complex algorithms and machine learning to train and adapt it to new information or situations. It goes beyond manually entering a value, as it can draw information from larger datasets, like the internet.

This ability to gather data from different places, use probabilistic models to weigh relevance to the task at hand, integrate that information, and then provide an output that uncannily resembles that of a human in many instances is what sets it apart from traditional computing. Large language models, like ChatGPT, showcase this new set of operations when you ask it a question and it cobbles together a specific answer. It takes the basic premise of a search engine but kicks it up a gear.

What concerns do you have about these changes to the nature of computation?

Lee: As AI products like ChatGPT and Bing become more popular, the nature of computing is becoming more inference based. This is a slight departure from the machine-learning models that were popular a few years ago, like the DeepMind's AlphaGO—the machine trained to be the best Go player—where the herculean effort was training the model and eventually demonstrating a novel capability. Now, massive AI models are being embedded into day-to-day operations like running a search, and that comes with trade-offs.

What are the material and resource costs associated with AI?

Jariwala: We take it for granted, but all the tasks our machines perform are transactions between memory and processors, and each of these transactions requires energy. As these tasks become more elaborate and data-intensive, two things begin to scale up exponentially: the need for

more memory storage and the need for more energy.

Regarding memory, an estimate from the Semiconductor Research Corporation, a consortium of all the major semiconductor companies, posits that if we continue to scale data at this rate, which is stored on memory made from silicon, we will outpace the global amount of silicon produced every year. So, pretty soon we will hit a wall where our silicon supply chains won't be able to keep up with the amount of data being generated.

Couple this with the fact that our computers currently consume roughly 20%–25% of the global energy supply, and we see another cause for concern. If we continue at this rate, by 2040 all the power we produce will be needed just for computing, further exacerbating the current energy crisis.

Lee: There is also concern about the operational carbon emissions from computation. So even before products like ChatGPT started getting a lot of attention, the rise of AI led to significant growth in data centers, facilities dedicated to housing IT infrastructure for data processing, management, and storage.

And companies like Amazon, Google, and Meta have been building more and more of these massive facilities all over the country. In fact, data center power and carbon emissions associated with data centers doubled between 2017 and 2020. Each facility consumes in the order of 20 megawatts up to 40 megawatts of power, and most of the time data centers are running at 100% utilization, meaning all the processors are being kept busy with some work. So, a 20-megawatt facility probably draws 20 megawatts fairly consistently—enough to power roughly 16,000 households—computing as much as it can to amortize the costs of the data center, its servers, and power delivery systems.

And then there's the embodied carbon footprint, which is associated with construction and manufacturing. This hearkens back to building new semiconductor foundries and packaging all the chips we'll need to produce to keep up with increasing compute demand. These processes in and of themselves are extremely energy-intensive, expensive and have a carbon impact at each step.

What role do these data centers play, and why are more of them needed?

Lee: Data centers offer economies of scale. In the past, a lot of businesses would build their own facilities, which meant they'd have to pay for construction, IT equipment, server room management, etc. So nowadays, it's much easier to just "rent" space from Amazon Web Services. It's why cloud computing has taken off in the last decade.

And in recent years, the general-purpose processors that have been prevalent in data centers since the early '90s started being supplanted by specialized processors to meet the demands of modern computing.

Why is that, and how have computer architects responded to this constraint?

Lee: Tying back to scaling, two observations have had profound effects on computer processor architecture: Moore's law and Dennard scaling.

Moore's law states that the number of transistors on a chip—the parts that control the flow of electrons on a semiconductor material—doubles every two or so years and has historically set the cadence for developing smaller, faster chips. And Dennard's scaling suggests that doubling the number of transistors effectively means shrinking them but also maintaining their power density, so smaller chips meant more energy-

efficient chips.

In the last decade, these effects have started to slow down for several reasons related to the physical limits of the materials we use. This waning effect put the onus on architects to develop new ways to stay at the bleeding edge.

General-purpose processors just weren't fast enough at running several complex calculations at the same time, so computer architects started looking at alternative designs, which is why graphics processing units (GPUs) got a second look.

GPUs are particularly good at doing the sort of complex calculations essential for machine learning algorithms. These tend to be more linear algebra centric, like multiplying large matrices and adding complex vectors, so this has also significantly changed the landscape of computer architecture because they led to the creation of what we call domain-specific accelerators, pieces of hardware tailored to a particular application.

Accelerators are much more energy efficient because they're custom-made for a specific type of computer and also provide much better performance. So modern data centers are far more diverse than what you would have had 10 to 15 years ago. However, with that diversity comes new costs because we need new engineers to build and design these custom pieces of hardware.

What other hardware changes are we likely to see to accommodate new systems?

Jariwala: As I mentioned, each computational task is a transaction between memory and processing that requires some energy, so our lab,

in conjunction with Troy Olsson's lab, is trying to figure out ways to make each operation use fewer watts of power. One way to reduce this metric is through tightly integrating memory and processing units because these currently exist in two separate locations that are millimeters to centimeters apart so electricity needs to travel great distances to facilitate computation which makes it energy and time inefficient.

It's a bit like making a high-rise mall, where you save space and energy and reduce travel time by allowing people to use the elevators instead of having them walk to different locations like they would in a single-story strip mall. We call it vertically heterogenous-integrated architecture, and developing this is key to reducing energy consumption.

But effectively integrating memory and processing comes with its own challenges because they do inherently different things that you wouldn't want interfering with one another. So, these are the problems people like my colleagues and me aim to work around. We're trying to look for new types of materials that can facilitate designs for making energy-efficient memory devices that we can stack onto processors.

Do you have any closing thoughts?

Jariwala: By now, it should be clear that we have an 800-pound gorilla in the room; our computers and other devices are becoming insatiable energy beasts that we continue to feed. That's not to say AI and advancing it needs to stop because it's incredibly useful for important applications like accelerating the discovery of therapeutics. We just need to remain cognizant of the effects and keep pushing for more sustainable approaches to design, manufacturing, and consumption.

Provided by University of Pennsylvania