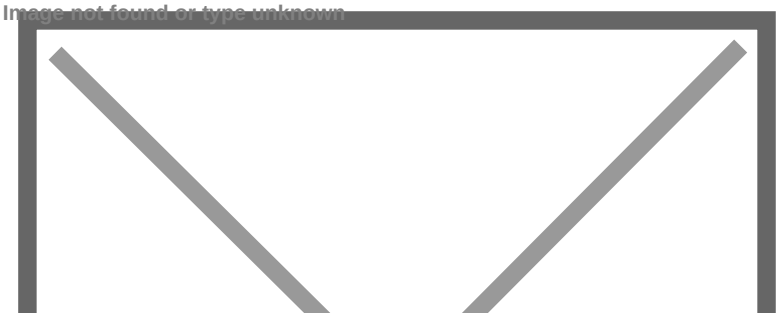
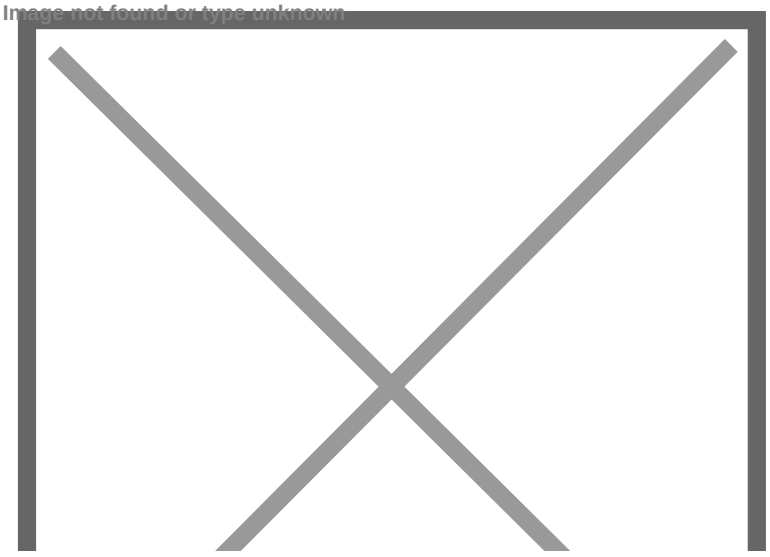
101: Ecosystem of AI



The massive ecosystem of AI relies on many kinds of extraction: from harvesting the data from our daily activities and expressions, to depleting natural resources and to exploiting labor around the globe so that this vast planetary network can be built and maintained.

This guide gives you insights, numbers and examples of art projects to help you as a maker navigate this field.



Cartography of Generative AI shows what set of extractions, agencies and resources allow us to converse online with a text-generating tool or to obtain images in a matter of seconds, by Estampa

“Cutting-edge technology doesn’t have to harm the planet”

Experimenting with AI

As a maker it can be daunting to experiment in this field, some tips (*this is a growing list*):

1. Be picky

Using large generative models to create outputs is far more energy intensive than using smaller AI models tailored for specific tasks. For example, using a generative model to classify movie reviews according to whether they are positive or negative consumes around 30 times more energy than using a fine-tuned model created specifically for that task.

The reason generative AI models use much more energy is that they are trying to do many things at once, such as generate, classify, and summarize text, instead of just one task, such as classification.

Be choosier about when they use generative AI and opt for more specialized, less carbon-intensive models where possible.

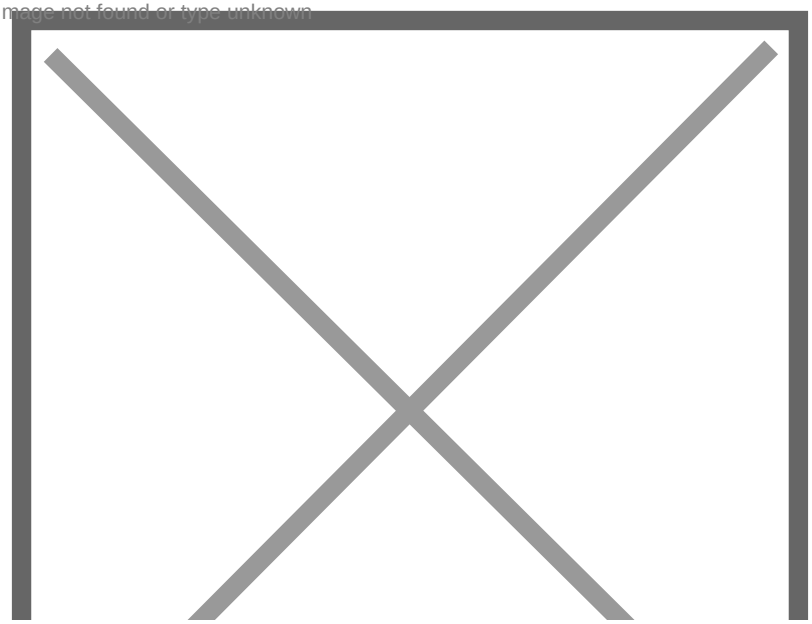
2. Use tools to keep track

Code Carbon makes these calculations by looking at the energy the computer consumes while running the model: https://codecarbon.io/

3. Run your models locally

Running AI models locally helps you to have increased control over energy usage and resource allocation. By managing models on personal or dedicated hardware, you can optimize efficiency and reduce energy consumption compared to relying on large, centralized data centers. This localized approach also minimizes the environmental impact, as it circumvents the need for extensive data transmission and the associated carbon emissions. Additionally, local operation enhances privacy and security, allowing sensitive data to remain on-site rather than being transmitted over potentially insecure networks. Read more in this book about how to do that.

Image not found or type unknown



Inspiration project: Solar Server is a solar-powered web server set up on the apartment balcony of Kara Stone to host low-carbon videogames. https://www.solarserver.games/

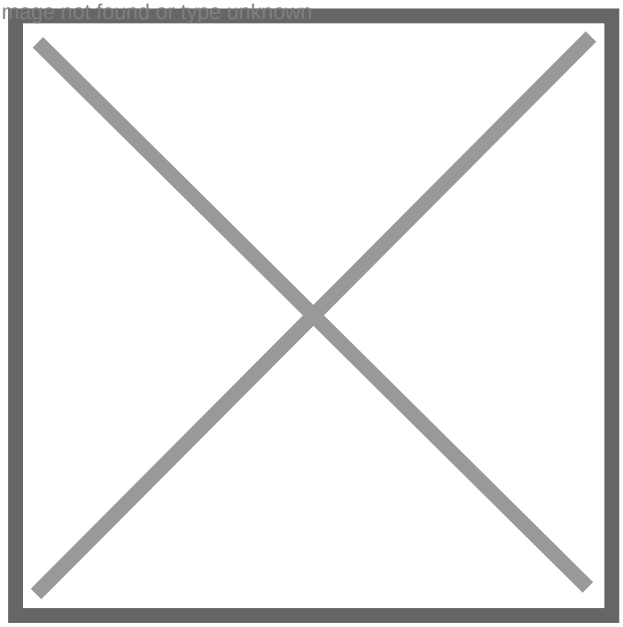
At what costs...

Artificial intelligence may invoke ideas of algorithms, data and cloud architectures, but none of that can function without the minerals and resources that build computings core components. The mining that makes

AI is both literal and metaphorical. The new extractivism of data mining also encompasses and propels the old extractivism of traditional mining. The full stack supply chain of AI reaches into capital, labor, and Earth's resources - and from each demands enormous amounts.

Each time you use AI to generate an image, write an email, or ask a chatbot a question, it comes at a cost to the planet. The processing demands of training AI models are still an emerging area of investigation. The exact amount of energy consumption produced is unknown; that information is kept as highly guarded corporate secrets. A reason is that we don't have standardized ways of measuring the emissions AI is responsible for. But most of their carbon footprint comes from their actual use. What we know:

Image not found or type unknown



The AI Index tracks the generative AI boom, model costs, and responsible AI use. 15 Graphs That Explain the State of AI in 2024:

Usage

- Making an image with generative AI uses as much energy as charging your phone. Creating text 1,000 times only uses as much energy as 16% of a full smartphone charge.
- Generating 1,000 images with a powerful AI model, such as Stable Diffusion XL, is responsible for roughly as much carbon dioxide as driving the equivalent of 4.1 miles in an average gasoline-powered car.
- A search driven by generative AI uses four to five times the energy of a conventional web search. Google estimated that an average online search used 0.3 watt-hours of electricity, equivalent to driving 0.0003 miles in a car. Today, that number is higher, because Google has integrated generative AI models into its search.
- It took over 590 million uses OF Hugging Face's multilingual AI model BLOOM to reach the carbon cost of training its biggest model. For very popular models, such as ChatGPT, it could take just a couple of weeks for such a model's usage emissions to exceed its training emissions.
- According to some estimates, popular models such as ChatGPT have up to 10 million users a day, many of whom prompt the model more than once.

Build/Run/Host

- Running only a single non commercial natural language processing model produces more than 660.000 pounds of carbon dioxide emissions, the equivalent of 5 gas powered cars over their total lifetime (incl manufacturing), 125 round trips New York - Beijing. This is a minimum baseline and nothing like the commercial scale Apple and Amazon are scraping internet-wide datasets and feeding their own NLP systems.*(AI researcher Emma Strubell and her team tried to understand the carbon footprint of natural language processing in 2019)*
- ChatGPT, the chatbot created by OpenAI, is already consuming the energy of 33,000 homes. OpenAI estimates that since 2012 the amount of compute used to train a single AI model has increased by a factor of ten every year.
- Data centers are among the world's largest consumers of electricity. China’s data center industry draws 73 percent of its power from coal, emitting about 99 million tons of CO2 in 2018, and the electricity is expected to increase two-thirds by 2023.
- Roughly two weeks of training for GPT-3 consumed about 700,000 liters of freshwater. The global AI demand is projected by 2027 to account for 4.2-6.6 billion cubic meters of water withdrawal, which is more than the total annual water withdrawal of Denmark or half of the United Kingdom.
- Water consumption in the company's data centres has increased by more than 60% in the last four years, an increase that parallels the rise of generative AI.

In-depth breakdown

The dirty work is far removed from the companies and city dwellers who profit most. Like the mining sector and data centers that are far removed from major population hubs. This contributes to our sense of the cloud being out of sight and abstracted away, when in fact it is material affecting the environment and climate in ways that are far from being fully recognized and accounted for.

Compute Maximalism

In the AI field it’s standard to maximize computational cycles to improve performance, in accordance with a belief that bigger is better. The computational technique of brute force testing in AI training runs or systematically gathering more data and using more computational cycles until a better result is achieved, has driven a steep increase in energy consumption.
Due to developers repeatedly finding ways to use more chips in parallel, and being willing to pay the price of doing so. The tendency toward compute maximalism has profound ecological impacts.

The [Uncertain] Four Seasons is a global project that recomposed Vivaldi's 'The Four Seasons' using climate data for every orchestra in the world. https://the-uncertain-four-seasons.info/project

Consequences

Within years, large AI systems are likely to need as much energy as entire nations.
Some corporations are responding to growing alarm about the energy consumption of large scale computation, with Apple and Google claiming to be carbon neutral (meaning they offset their carbon emissions by purchasing credits) and Microsoft promising to become carbon negative by 2030.
In 2023, Montevideo residents suffering from water shortages staged a series of protests against plans to build a Google data centre. In the face of the controversy over high consumption, the PR teams of Microsoft, Meta, Amazon and Google have committed to being water positive by 2030, a commitment based on investments in closed-loop systems on the one hand, but also on the recovery of water from elsewhere to compensate for the inevitable consumption and evaporation that occurs in cooling systems.

Image not found or type unknown

Deep Down Tidal is a video essay by Tabita Rezaire weaving together cosmological, spiritual, political and technological narratives about water and its role in communication, then and now.

More about water

Reflecting on media and technology and geological processes enables us to consider the radical depletion of nonrenewable resources required to drive the technologies of the present moment. Each object in the extended network of an AI system, from network routers to batteries to data centers is built using elements that require billions of years to form inside the earth.

Water tells a story of computation's true cost. The geopolitics of water are deeply combined with the mechanisms and politics of data centers, computation, and power - in every sense.

The digital industry cannot function without generating heat. Digital content processing raises the temperature of the rooms that house server racks in data centres. To control the thermodynamic threat, data centres rely on air conditioning equipment that consumes more than 40% of the center's electricity (Weng et al., 2021). But this is not enough: as the additional power consumption required to adapt to AI generates more heat, data centers also need alternative cooling methods, such as liquid cooling systems. Servers are connected to pipes carrying cold water, which is pumped from large neighboring stations and fed back to water towers, which use large fans to dissipate the heat and suck in freshwater
The construction of new data centers puts pressure on local water resources and adds to the problems of water scarcity caused by climate change. Droughts affect groundwater levels in particularly water-stressed areas, and conflicts between local communities and the interests of the platforms are beginning to emerge.

Curious about more? Reading tips

Atlas of AI - Kate Crawford

A Geology of Media - Jussi Parikka

Hyper objects - Timothy Morton

Sources:

Atlas of AI - Kate Crawford

Podcast: Kunstmatig #28 - Tussen zeespiegel en smartphone

Datacenters and water: KUER

Technology Review: Making an Image with Gen AI uses as much energy as charging your phone

https://www.washingtonpost.com/technology/2023/06/05/chatgpt-hidden-cost-gpu-compute/

https://arxiv.org/pdf/2206.05229

Technology Review: getting a better idea of gen AIs footprint

https://spectrum.ieee.org/ai-index-2024

Standford EDU: Measuring trends in AI

Revision #7
Created 25 June 2024 09:29:22 by Cyanne
Updated 27 June 2024 07:23:10 by mikal