AI and the ENERGY TRANSITION

Summary of workshop held 11 March 2024 at the NTNU Energy Transition Week, Trondheim Norway

BRIEF

NTNU ENERGY TRANSITION

NTNU

Brief 01/2024

Link to workshop programme

CONTRIBUTORS:

Amparo Alonso-Betanzos, University of A Coruña and NTNU (main editor)
Astrid Sørensen, Equinor
Bjarne Foss, NTNU
Coral Calero, University of Castilla-La Mancha
Jorge Paz-Ruza, University of A Coruña
Brais Cancela Barizo, University of A Coruña
Oscar Fontenla-Romero, University of A Coruña
Bertha Guijarro-Berdiñas, University of A Coruña
Christian A. Klöckner, NTNU
Rialda Spahic, Equinor
Justin Fackrell, Equinor
Frauke Behrendt, Eindhoven University of Technology
Timothy Capper, OptiSpark
Nhan Van Nguyen, eSmart Systems
Pablo Fariñas Alvariño, University of A Coruña
José Mª Álvarez Gallego, Inditex
Gunnar Tufte, NTNU

Cite as:

Alonso-Betanzos, A (ed), A. Sørensen, B. Foss, C. Calero, J. Paz-Ruza, B. Cancela-Barizo, O. Fontenla-Romero, B. Guijarro-Berdiñas, C. Klöckner, R. Spahic, J. Fackrell. F. Berendt, T. Capper, N. Nguyen, P. Fariñas-Alvariño, J. Alvarez-Gallego, G. Tufte (2024): AI and the Energy Transition – workshop summary, NETI brief 01/2024, NTNU, Trondheim, Norway

Note:

This report should not be used as scientific facts and conclusion, but rather as a summary of important issues and aspects discussed at the workshop.

Print: Grafisk Senter

SUMMARY

In an era where artificial intelligence (AI) intersects with every facet of our lives, the dialogue around its sustainable integration becomes paramount. Our recent workshop on AI and Sustainability served as a vibrant platform for exploring this critical intersection. This document encapsulates the rich discussions, insightful keynotes, and thought-provoking interventions from leading experts across various domains. The aim of the workshop was to illustrate the pivotal role that AI plays in shaping a sustainable future. And we did so by exploring two crucial complementary aspects: Sustainable AI and AI for Sustainability.

As we stand at the crossroads of technological innovation and environmental responsibility, the insights gleaned from this workshop offer a guiding light. This summary not only reflects our shared learnings but also embodies our commitment to fostering an AI-enabled future that harmonizes with our planet's ecological boundaries.

Part I: Sustainable AI

Our first segment explored the concept of Sustainable AI, emphasizing the need to make AI development environmentally conscious and ethically responsible. AI, with its ever-growing computational demands driven by increasingly complex algorithms, poses a significant challenge and brings to light the pressing need for AI systems designed with environmental conscientiousness at their core. The rapid escalation in AI's computational appetite, fueled by intricate algorithms, poses substantial challenges. For instance, the training of GPT-3 consumed 1.300.000 kWh, equivalent to 552 tons CO_2 emissions, and 700.000 liters of water for data center refrigeration, equivalent to the annual energy consumption of 126 households. Predictions even suggest a potential increase, possibly reaching 30% of the world's consumption.

In Figure 1, it can be seen that training recent models requires important computational resources and has an equivalent in CO_2 emissions. The columns in blue reflect the emissions of the models, while the violet show real-life examples for comparison. In parenthesis, beside the name of the models, it is shown the number of parameters of the models to be adjusted. It can be seen also that there is not always a direct relation between that number and the emissions (see for example GPT-3 and OPT, both with 175 billion parameters, but very different CO_2 emissions).

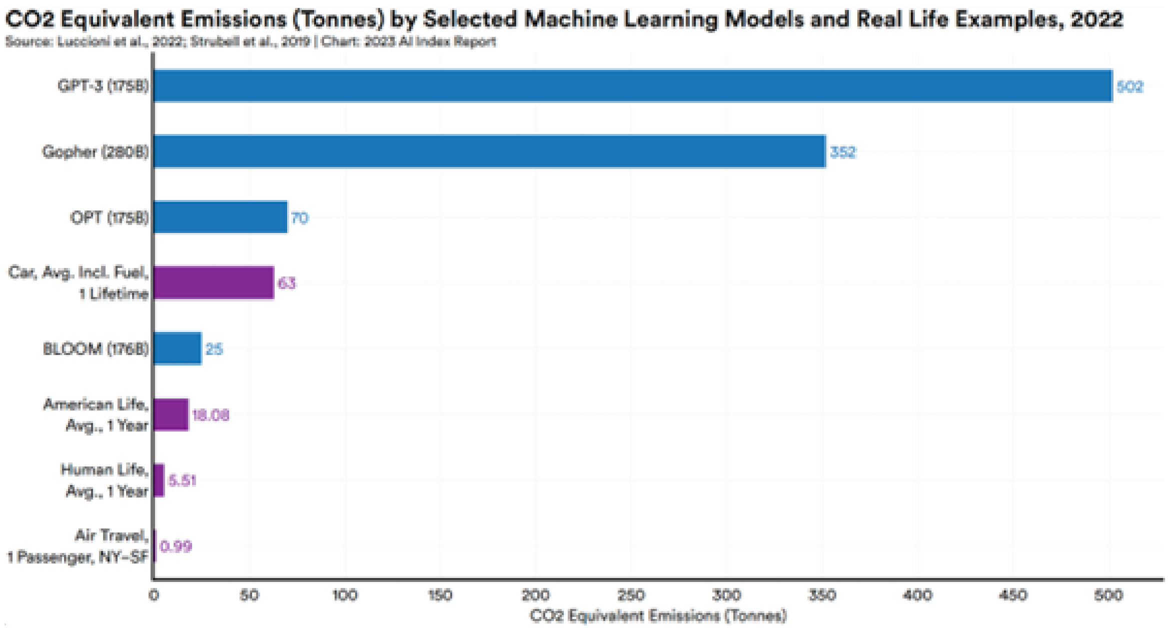


Figure 1: CO_2 equivalent emissions (Tonnes) by selected machine learning models and real life examples, 2022 (Source: Luccioni et al., 2022, Strubell et al., 2019, Chart: 2023 AI Index Report).

Sustainable AI not only seeks to minimize the ecological footprint of AI technologies but also emphasizes transparency, fairness, and inclusivity in their design and deployment. Throughout this part of the workshop, we explored how AI can positively contribute to ecological well-being. Different perspectives on conscious use of tools, calculators for energy consumption of our algorithms and several ideas for the so-called frugal AI were enumerated, while the perspective of the efficiency of the human brain for intelligent tasks and the need to continue pursuing ideas on how to translate the natural mechanisms of our brain into AI models and tools were highlighted.

Part II: AI for Sustainability

Moving forward, the second part of our workshop focused on leveraging AI as a powerful tool for addressing and mitigating sustainability challenges, acting as a formidable ally in the global quest for sustainability.

The workshop unfolded against the backdrop of our finite planetary resources; a theme poignantly introduced by Professor Gunnar Tufte from NTNU.

AI for Sustainability explores how innovative applications of artificial intelligence can enhance our ability to monitor, analyze, and solve complex environmental and social issues. Our discussions spanned the spectrum of AI's impact, touching on its integration into city planning, urban life, and industry-specific challenges such as predictive maintenance and supply chain optimization. Notably, the dialogue ventured beyond the technical intricacies of AI, delving into the psycho-sociological implications and the pressing need for regulatory frameworks to guide its ethical deployment.

Governance for sustainable and just use of AI in urban life e.g., will seek answers to new important questions when shaping sustainable futures and decision making. Examples here are; How to best care for sustainability in hybrid governance processes?; How to translate policy goals, such as emission reduction or accessibility into AI?; How to co-create strategies for hybrid governance that allow for open trade-offs between individual, market, and public interests – especially in the light of climate crisis?; and Who should be involved an how: developers, policymakers, company leaders, civil society? And what skills is required? (Frauke Behrendt, TU/e).

A recurrent theme was the balance between technological advancement and environmental stewardship. The workshop underscored the imperative to harness AI not merely as a tool for efficiency but as a catalyst for sustainable innovation. As we navigated the complexities of Moore's law, energy consumption, and the limitations imposed by physical constraints, the dialogue converged on the critical need for a responsible approach to AI development.

This workshop summary continues with a closer look at the workshop’s aspect of Sustainable AI before ending with inspiration from neuroscience.

SUSTAINABLE AI

During the last years, the accuracy of the modern AI models increased, but at the expense of an important increase in the complexity of those models, that in turn reflects in crescent energy consumptions, and their equivalent CO_2 emissions (Figure 1). This is the so-called "red AI", driven almost exclusively by accuracy.

However, for AI to effectively support an appropriate energy transition, the initial step is to develop new models and algorithms that incorporate a multidimensional assessment. This entails considering environmental impacts, alongside social and economic factors, to ensure a balanced approach.

How to measure energy consumption of algorithms

But, how to measure the energy consumption of algorithms, so we can compare several alternatives? Thus, our first quest revolves around quantifying the environmental footprint of software and enhancing its energy efficiency, all the while keeping a keen eye on the socio-economic dimensions intertwined with software sustainability. To put this into perspective, consider the energy footprint of streaming services like Netflix. An hour of streaming equates to the carbon emissions from driving a car 200 meters. While seemingly insignificant in isolation, this scales dramatically when we account for the global consumption rates, painting a stark picture of the environmental impact.

The imperative to mitigate software's environmental footprint is clear, echoing Tom DeMarco's sentiment that control is predicated on our ability to measure. Our exploration into sustainable AI practices is two-fold, encompassing both software- based and hardware-based approaches to energy measurement. Each pathway, with its distinct methodologies and tools, offers insights into the intricate dance between technological innovation and environmental stewardship.

Then, the workshop explored different perspectives and models for a more sustainable AI, such as:

Reducing the footprint of Deep Learning (DL) Models

DL is one of the recent machine learning models that has led to significant advancements in various fields, such as computer vision, natural language processing or speech recognition. They consist of multiple neural networks of interconnected nodes, and are often computationally intensive, requiring large amounts of data and processing power to train efficiently. The training process involves iterative computations and adjustments to millions or even billions of parameters, leading to significant energy consumption, which can contribute to environmental concerns, including carbon emissions.

To mitigate the energy consumption associated with deep learning, researchers are exploring various approaches, including optimizing algorithms for efficiency, developing hardware accelerators tailored for deep learning tasks, and exploring alternative training methods such as transfer learning and self-supervised learning. Additionally, efforts are underway to develop more energy-efficient training techniques and to increase the utilization of renewable energy sources in data centers to reduce the environmental impact of deep learning applications. There are also options available for reducing the carbon footprint for the models already in inference, such as:

- **Quantization**, which involves honing the precision of data in model training, transitioning from high-bit values to more compact formats like integers or 8-bit representations. The magic of quantization lies in its potential to significantly reduce model size—up to 16-fold—while meticulously preserving accuracy, often achievable with an eightfold reduction. This balance is crucial in maintaining the integrity of model outputs while embracing energy efficiency.

- **Distillation**, which involves the transfer of knowledge from a larger 'teacher' model to a more streamlined 'student' model. The aim here is to cultivate a student model that mirrors the teacher's predictive prowess, often surpassing the performance of directly training the student model on the dataset.

- **Pruning**, in which strategic elimination of selected nodes and channels within deep learning architectures, refining their structure and functionality is carried out. While the primary goal is to enhance accuracy, pruning also holds the potential to diminish the storage demands of the model. However, its contribution to energy savings is somewhat constrained, given the necessity to load the complete model into memory, especially when dealing with sparse matrices.

Beyond these more conventional techniques, there are also other approaches such as low-rank quantization, sparse representations, and feature-matched compression.

Long-life learning models

Long-life learning models refer to a category of machine learning techniques designed to enable AI systems to continuously learn and adapt over time from new data and experiences without catastrophically forgetting previously learned knowledge. These models address the challenge of retaining and leveraging past knowledge while accommodating new information, which is crucial for tasks requiring ongoing adaptation to changing environments or requirements. This adaptive learning approach not only aligns with our innate learning mechanisms but also presents a promising avenue for reducing the environmental footprint of AI systems.

Key strategies in long-life learning models include leveraging techniques such as elastic weight consolidation (EWC), replay buffers, and dynamic architectures to manage the trade-off between retaining old knowledge and accommodating new information. By enabling AI systems to learn incrementally and adaptively, long-life learning models aim to facilitate the development of more flexible, robust, and capable AI systems for real-world applications. Furthermore, by fostering models that build upon existing knowledge, we can significantly reduce the need for resource-intensive training from scratch, thereby mitigating both energy and water consumption in AI operations. By embracing these methodologies, we can pave the way for AI systems that are not only intelligent and versatile but also environmentally conscious and sustainable.

Federated Learning and Edge computing

Traditional machine learning uses training data that is centralized, but the rise of distributed data generation — be it across different institutions or an array of IoT devices—presents new challenges and opportunities for decentralization. An available and more environmental-friendly option is federated learning, a collaborative framework where diverse devices contribute to developing a comprehensive model, each leveraging its local data. After local training, these partial models converge through a coordinating device, culminating in a unified global model. This model, once refined, is redistributed to the participant devices for application on new data. Besides, federated learning epitomizes the collaborative spirit of AI, fostering models that respect privacy and computational constraints.

Personalized and greener models

At its core, personalization tailors AI-driven processes to align with the unique preferences, needs, or expectations of individual users. This aspect, intrinsic to the digital experience, underscores the indispensability of personalization across various AI applications, transcending mere convenience to embody a deeper connection between technology and its human counterparts.

Personalization's spectrum extends far beyond targeted advertisements, encapsulating domains like personalized medicine, where treatment regimens are fine-tuned to individual patient profiles, or the automotive industry's adaptive driving modes, attuned to the nuances of a driver's habits and prevailing road conditions. Such applications not only enhance user engagement but also foster a more equitable and trustworthy interaction with AI systems.

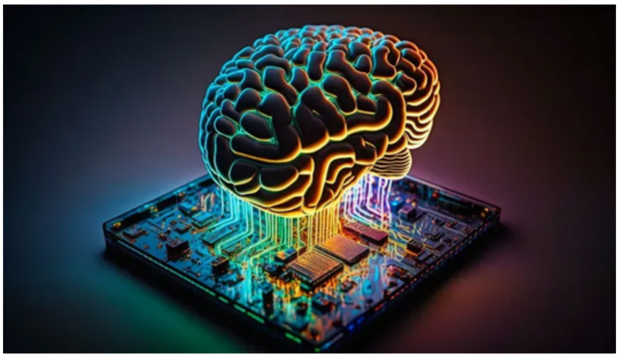
However, personalization usually comes with a cost — financially and ecologically. Tailoring AI to individual users necessitates a granular understanding of each user, thereby amplifying the model's computational demands. Thus, it becomes imperative to scrutinize and steer our policies and research endeavors towards sustainability in AI personalization.

Personalization manifests in two primary forms: personalized outputs, where the AI's output varies to cater to individual users, and the personalization of presentation, aimed at enhancing interpretability and user trust. Both avenues, while enriching the user experience, beckon us to weigh their environmental implications carefully. However, it is possible to minimize the footprint of the models while performance is maintained or even increased, a relevant question that involves both the use of quality data as well as a thoughtful model design.

And last, but not least:

Inspiration from neuroscience; Neuromorphic computing

One option for sustainability involves seeking inspiration from Neuroscience. The human brain operates with remarkable energy efficiency, utilizing approximately 20 watts of power— a stark contrast to the high energy demands of contemporary AI systems. Besides, the brain's proficiency in learning from unique experiences, its resilience to noise, and its adept handling of multimodal inputs are attributes that AI systems aspire to emulate.



Neuromorphic computing (an approach to artificial intelligence that seeks to mimic the architecture and function of the human brain) is a concept introduced by Carver Mead, and encapsulates this aspiration towards brain-inspired computing, offering a promising pathway for AI's evolution. There are several relevant advantages of neuromorphic computing that are to be taken into account for different aspects of sustainability:

- **Real-life learning and adaptability,** as these systems can learn and adjust with few data, responding quickly to changing environments and situations, overcoming the extensive training data requirements of most present AI systems.

- **Energy-efficiency**, by leveraging electronic circuits inspired by biological neurons. This could allow also for working in resource-limited settings.

- **Resilience to noise and error**, as information is processed through distributed networks of neurons.

- **Privacy and security**, processing data locally, and ensuring privacy for sensitive information.

Through research in the area, deeper insights into AI operation might be gained, leading to more transparent and explainable systems.

Amparo Alonso Betanzos
Professor at the University of Coruna and NTNU
Head of workshop organizing committee

NTNU ENERGY TRANSITION

NTNU