Frugal AI: Introduction, Concepts, Development and Open Questions

Ludovic Arga, François Bélorgey, Arnaud Braud, Romain Carbou, Nathalie Charbonniaud, Catherine Colomes, Lionel Delphin-Poulat, David Excoffier, Christel Fauché, Thomas George, Frédéric Guyard, Thomas Hassan, Quentin Lampin, Vincent Lemaire, Pierre Nodet, Pawel Piotrowski, Krzysztof Sapiejewski, Emilie Sirvent-Hien, Tamara Tosic

(authors alphabetical order, contact: firstname.name@orange.com)

Orange Research

ABSTRACT

This document aims to provide an overview and synopsis of frugal AI, with a particular focus on its role in promoting cost-effective and sustainable innovation in the context of limited resources. It discusses the environmental impact of AI technologies and the importance of optimising AI systems for efficiency and accessibility. It explains the interface between AI, sustainability and innovation. In fourteen sections, it also makes interested readers aware of various research topics related to frugal AI, raises open questions for further exploration, and provides pointers and references.

1. INTRODUCTION ABOUT THIS DOCUMENT

About this document - The objective of this document is to provide a preliminary synopsis of frugal AI, with a particular emphasis on its role in fostering cost-effective and sustainable innovation in the context of limited resources. It discusses the environmental impact of AI technologies and the importance of optimising AI systems for efficiency and accessibility. The authors do not pretend to cover all the aspects of frugal AI but give understanding in the intersection of AI, sustainability, and innovation. The document aims to raise awareness of interested readers about various related topics, poses open questions for further exploration in the field of frugal AI, provides some pointers and references. The different sections have been written independently so that the reader can read only one part without reading the full document. As a result, there is potential redundancy between some of the sections presented.

Frugal AI is at the intersection of 4 domains: the economy, the technology, the society, and climate change. The figure 1 below introduces, as a snapshot, the main concepts that will be detailed in the document.

The document delineates the notion of frugal AI, highlighting its capacity for cost-effective and sustainable innovation in resource-constrained environments. It emphasises the environmental impact of AI technologies and the necessity for optimising AI systems to reduce their ecological footprint. The document goes on to explore a variety of strategies for achieving frugality in AI, including the right usage of AI, model compression, hardware optimization, and the importance of resource-aware AI design. The document also poses a series of research questions to stimulate further investigation into the implications of frugal AI across economic, social, and environmental domains.

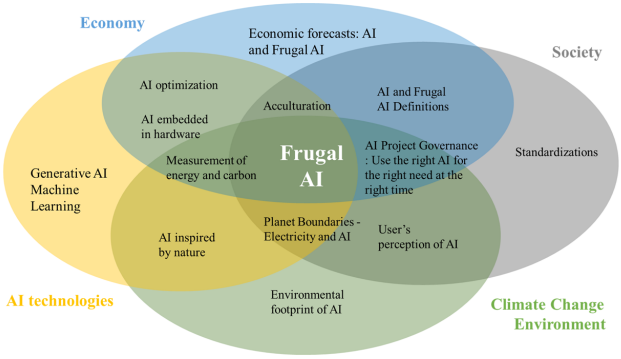

Figure 1: Frugal AI: what impact and what solutions for the environment?

Introduction - Generative AI, the latest digital revolution, is transforming the way we use digital technology in our daily lives and is also highlighting fundamental issues such as responsibility, safety, and ethics. However, the environmental footprint of digital technology is often overlooked, even though it already accounts for nearly 4% of global greenhouse gas emissions. With the rise of artificial intelligence, this footprint will increase, putting pressure on vital resources such as electricity and water in certain regions of the world. Faced with this dilemma, the concept of frugal AI is emerging. It questions the tension between the unbridled development of artificial intelligence and the planetary limits we keep pushing.

In the past, frugal innovation was a strategy suitable only for low-income countries where there were severe resource constraints. However, raising barriers to recent innovation

thinking makes frugal innovation best suited to all levels of a nation's development. The word "frugal" is a well-known definition for being thrifty or economical. But when frugal modifies innovation, its acronym should be parsed as follows: functional, robust, user-friendly, growing, affordable, and local [88].

Frugal Innovation is an opportunity to innovate cost-effectively and sustainably under resource scarcity. Like the poet Charles Baudelaire, who said of poetry that "*because the form is constrained, the idea springs forth more intensely*", we propose to ask ourselves: what innovative ideas are emerging or will emerge from these constraints?

Outline of this document - Below is a brief roadmap outlining the contents and contributions of each section:

- Section 2 defines AI systems and their relevance in frugal innovation. It clarifies what constitutes an AI system and defines frugal AI, highlighting its focus on efficiency and resource-conscious solutions. It also explains how AI can be utilized to create cost-effective solutions in resource-constrained settings.
- Section 3 discusses the environmental consequences of AI technologies, including energy consumption, greenhouse gas emissions and resource consumption. It examines the ecological impact of generative AI, its resource demands and the implications of generative AI.
- Section 4 analyzes public perceptions and awareness of AI, particularly generative AI, and its societal implications.
- Section 5 explores the economic landscape for AI and frugal AI, including investment trends and labour market implications.
- Section 6 discusses the challenges posed by energy and resource limitations on AI growth.
- Section 7 highlights the importance of selecting appropriate AI models based on performance and resource efficiency.
- Section 8 introduces methods for measuring the environmental impacts of AI throughout its lifecycle.
- Section 9 stresses the need for public education on the environmental impacts of AI and the principles of frugal AI.
- Section 10 discusses the importance of establishing standards for the design and deployment of AI to minimize environmental impacts.
- Section 11 explores how natural systems can inspire the design of frugal AI solutions.
- Section 12 examines advancements in hardware that support frugal AI applications, focusing on energy efficiency and cost-effectiveness.
- Section 13 reviews various strategies for optimizing AI models, including model compression and hardware optimization.
- Section 14 presents open questions and areas for further exploration in the field of frugal AI.

This roadmap provides a structured overview of the manuscript, allowing readers to quickly grasp the key themes and contributions of each section related to frugal AI.

2. CONTEXT AND DEFINITION

2.1 What is an AI system

Understanding artificial intelligence is the first step towards understanding the concept of frugal AI. Here are the main definitions:

- **AI system** [EU AI Act Article 3[1]] is a machine-based framework with varying levels of autonomy. It can be adapted to achieve explicit or implicit goals. The system processes the current input data and produces the outcome result (i.e., detection, prediction, content generation or recommendation) that can influence physical or virtual environments.

- **AI Expert system** [ISO/IEC 22989[2]]: AI system that accumulates, combines and encapsulates knowledge provided by a human expert or experts in a specific domain to infer solutions to problems.

- As AI is widely used in social debates, the **social definition of AI** today takes an important place. Will Heaven in the MIT Technology Review defines it as a catch-all word: *"AI is a catch-all term for a set of technologies that make computers do things that are thought to require intelligence when done by people."* [98]. We can follow Hubert Guillaud in his attempt to define AI as the set of techniques that stand between lab research and widespread usages [91]. In 2025, AI is identified as LLMs (Large Language Models), but before this, AI was used to describe image recognition. Beyond the ambiguity of the term that covers both a field of computer science, but also techniques that articulate models on data, Alex Bender and Emily Hanna also point out that artificial intelligence comes with "magic" and could be omniscient and all-powerful [93]. In addition, the human anthropomorphizes the machine, that is to say, attributes it an intention [60]. This is especially the case for AI methods that use language models. It is therefore important to educate populations to keep critical thinking in AI usages, to avoid replacing prompts for questions and feedback for answers.

2.2 Defining AI in the context of frugality

Artificial intelligence (AI) in the context of frugal innovation refers to the use of intelligent technologies to develop cost-effective, efficient, and resource-conscious solutions. AI enables systems to learn from data, automate processes, and make informed decisions, often with minimal human intervention. In frugal innovation, AI is applied to create solutions that are accessible, affordable, and adaptable to resource-constrained environments. In essence, frugal innovation seeks to develop high-value solutions using minimal resources.

By leveraging optimization techniques, AI can function effectively within the constraints of limited infrastructure, making it an indispensable tool in contexts where conventional

[1] `https://artificialintelligenceact.eu/article/3/`
[2] `https://www.iso.org/fr/standard/74296.html`

approaches may be impractical. As [88] highlights, frugal innovation can be significantly enhanced by technological advancements. Citing [239], Govindan asserts that AI holds a distinct advantage over other technologies in fostering frugal innovation. Additionally, Govindan references [224]'s argument that AI-driven improvements in frugal innovation can contribute to a company's growth. These perspectives support the central question explored in Govindan's research: *What is the significance of integrating AI into sustainable frugal innovation?*

Despite its potential, the integration of AI into sustainable frugal innovation presents several challenges. Entrepreneurs and organizations often face difficulties in aligning AI-driven solutions with sustainable innovation strategies. As noted by [88], understanding the **critical success factors** (CSFs) for AI implementation is essential for overcoming these barriers. This paper raises two fundamental questions: *What are the common drivers for AI implementation in sustainable frugal innovation?* and *Which of these factors exert the most significant influence?*

Govindan's study identifies "understanding the concept of AI" and "level of AI investment" as the two most influential success factors for AI adoption in sustainable frugal innovation [88]. These factors are critical in determining how industries can integrate AI-driven solutions to enhance their business competitiveness, particularly in times of disruption [88]. The study suggests that by addressing these key factors, businesses can maximize AI's potential in fostering cost-effective, scalable, and sustainable innovation.

To ensure the successful integration of AI into sustainable frugal innovation, [88] emphasizes the need for targeted strategies aimed at strengthening these key success factors. The study highlights that industries must develop specific practices to facilitate AI adoption. One of the most effective approaches, according to [88], is providing structured training for employees and top-level management. This can be achieved through participation in workshops and seminars, as well as engaging with technical literature on AI applications in sustainable frugal innovation. Such initiatives enhance decision-making by improving organizational understanding of AI's role in resource-efficient innovation.

By fostering AI literacy and ensuring strategic investments, industries can unlock the full potential of AI-driven frugal innovation. As [88] suggests, a well-informed approach to AI integration can contribute to long-term sustainability and resilience, enabling businesses to thrive in increasingly resource-conscious environments. The ongoing exploration of AI's role in frugal innovation will therefore remain critical for industries seeking to maintain competitiveness while addressing global sustainability challenges.

2.2.1 Frugal Artificial Intelligence (FAI)

Artificial Intelligence (AI) has become increasingly sophisticated, with machine learning (ML) models achieving higher accuracy in various applications. However, this progress often comes at a significant computational and environmental cost. The development and deployment of AI models require extensive data preprocessing, substantial computing resources, considerable energy consumption, and in consequence, CO_2 footprint of the training process, raising concerns about sustainability and accessibility [125]. In response to these challenges, the concept of Frugal Artificial Intelligence (FAI) has emerged as a framework aimed at reducing AI's resource dependency while maintaining its effectiveness. As [125] stated, "Here, frugality can concern (this list is not exhaustive):

1. Reduction of data size, i.e., minimization of dataset(s) used in training, while preserving model accuracy.

2. Making AI eco-friendly, by reducing the energy involved in model training and use.

3. Minimization of needed resources, i.e., memory and/or processing/battery power"

2.2.2 Key principles of FAI

1. Efficiency: Frugal AI solutions prioritize efficiency in terms of both computation and energy consumption. This may involve designing algorithms that can run on inexpensive hardware or optimizing code to minimize resource usage.

2. Affordability: Frugal AI aims to make AI technologies accessible to a wide range of users, regardless of their financial resources. This may involve reducing the cost of hardware, software, and infrastructure required for AI implementation.

3. Simplicity: Frugal AI solutions often prioritize simplicity and ease of use over complexity. This may involve using simpler algorithms or user interfaces that require less training and technical expertise to operate.

4. Scalability: Frugal AI solutions should be scalable, allowing them to adapt to different contexts and user needs without significantly increasing costs. This may involve designing modular architectures that can be easily expanded or customized as needed.

2.2.3 Ways to make AI frugal

To build frugal AI methods by design, as a society, we should consider these key points, discussed in more detail in the next sections:

- understand the impact that AI has on our planet and society (see Sections 3, 6, 9),
- apply eco-design of AI (see Section 8),
- understand the alternative setups with limited resources (see Sections 12, 13, and 11),
- conceive our AI for current needs and usages (final training model and its intermediate steps),
- apply recommendations, specifications and regulations (see Sections 10, 7).

2.3 Frugality versus efficiency in the context of artificial intelligence

Artificial intelligence (AI) has evolved through various paradigms, each offering distinct approaches to solving problems. As AI technologies advance, two key concepts - frugality and efficiency - have emerged as critical considerations in both research and practical applications. Although these terms may seem similar, they encapsulate different principles in the design and deployment of AI systems. In this chapter, we explore these differences in detail.

2.3.1 Defining Efficiency in AI

In the context of AI, efficiency generally refers to the optimal use of resources to achieve a specific performance goal. Key aspects include:

1. Computational Efficiency: This involves minimizing the amount of time, memory, or energy required to execute an algorithm. Efficient AI systems perform tasks faster and with fewer computational resources.

2. Algorithmic Efficiency: Here, the focus is on designing algorithms that achieve high accuracy and performance while operating within acceptable resource limits. For example, an efficient algorithm might deliver accuracy similar to that of a more complex one but with lower computational costs.

3. Operational Efficiency: This can include aspects such as scalability (the ability to handle increasing amounts of work) and cost-effectiveness during deployment. In many cases, efficiency improvements are measured by the trade-off between output quality and resource input.

In summary, efficiency in AI is largely about optimization - making sure that every computational resource (whether it be time, energy, or memory) is used to its fullest potential to achieve the desired outcomes.

2.3.2 Understanding Frugality in AI

While efficiency focuses on optimal resource utilization, frugality embodies a broader philosophy. It goes beyond mere optimization to encompass the design of AI systems that are inherently resource-conscious from the outset. Key characteristics of frugality include:

1. Minimalism in Design: Frugal AI systems are built with the principle of "less is more." This means they are designed to function effectively with minimal resources, avoiding unnecessary complexity.

2. Accessibility and Affordability: Frugality emphasizes creating AI solutions that are accessible in resource-constrained environments. This is particularly important for applications in developing regions or for organizations with limited budgets.

3. Sustainable Innovation: Frugal AI takes into account long-term sustainability. It aims to reduce environmental impacts by minimizing energy consumption and promoting the use of available resources wisely.

4. Context-Aware Development: In frugal innovation, the design process begins with a clear understanding of the specific resource constraints and needs of the target environment. This can lead to novel, context-specific approaches that differ from traditional, resource-intensive AI models.

Thus, while efficiency is about optimizing existing processes, frugality is a proactive strategy. It involves designing full AI systems to operate under strict resource constraints, often resulting in solutions that are both cost-effective and sustainable.

2.3.3 Terminology: Frugality, Efficiency, and Related Concepts

In the literature, several terms are used interchangeably to describe aspects of resource management in AI. Understanding these terms can help clarify the distinction between frugality and efficiency:

1. Lean AI: Borrowed from lean manufacturing principles, lean AI emphasizes minimizing waste and unnecessary complexity. This concept aligns closely with frugality, as it promotes the development of streamlined, purpose-built systems.

2. Sustainable AI: Sustainable AI focuses on reducing the environmental footprint of AI systems, including energy consumption and electronic waste. This concept is an important aspect of frugality, though it also overlaps with efficiency when considering operational costs.

3. Green AI*: The term Green AI [194] refers[3] to AI research that yields novel results without increasing computational cost, and ideally reducing it. Whereas Red AI has resulted in rapidly escalating computational (and thus carbon) costs, Green AI has the opposite effect. If measures of efficiency are widely accepted as important evaluation metrics for research alongside accuracy, then researchers will have the option of focusing on the efficiency of their models with a positive impact on both the environment and inclusiveness.

4. Responsible AI*: Responsible Artificial Intelligence (Responsible AI) is an approach[4] to developing, assessing, and deploying AI systems in a safe, trustworthy, and ethical way and promoting positive outcome.

* Note: These terms are very commonly used, although they are not really defined in the standards.

2.3.4 Distinguishing Frugality from Efficiency in AI

While these terms share common ground, they differ in scope and emphasis. They represent different approaches:

1. Focus and Intent:

    - **Efficiency** focuses on optimizing performance metrics (such as speed, accuracy, and energy usage) within a given framework. The goal is to maximize output for any fixed level of resource input.

[3]Subsequently, the term has evolved in meaning and sometimes also refers to AIs designed to optimise environmental impact.

[4]Sometimes positioned differently in French (the right solution for the right need) mainly because of the difference in meaning of the word "responsible" in English and "responsible" in French.

- **Frugality** emphasizes a minimalistic design philosophy. It starts with the assumption that resources are scarce and seeks to develop solutions that are inherently low-cost and sustainable, rather than simply optimizing existing processes.

2. Design Versus Optimization:

   - **Efficiency** improvements are often applied as optimizations to existing systems, such as refining algorithms or reducing computational overhead.
   - **Frugal innovation** involves rethinking the system from the ground up, incorporating resource constraints into the design process itself. This can lead to entirely new approaches that differ from traditional methods.

3. Context and Application:

   - **Efficiency** is a universal goal across many fields of AI, regardless of the operating environment.
   - **Frugality** is particularly relevant in contexts where resource limitations are a fundamental constraint, such as in developing regions or in applications with strict energy budgets. Frugal AI is not just about doing more with less, but about designing accessible and sustainable methods over the long term.

In essence, while both concepts value resource conservation, efficiency is about doing things better, and frugality is about doing things differently, with a focus on simplicity, accessibility, and sustainability.

3. WHAT IS THE ENVIRONMENTAL FOOTPRINT OF AI

In 2023, greenhouse Gas (GHG) emissions due to the digital domain represented nearly 4% of the global GHG emissions. Shortly, this contribution will be doubled due to IA expansion. One knows that AI is water and power-greedy at least, which gives it a major role in the GHG emissions increase of the digital sector. Here is an overview of the environmental impact of IA.

3.1 Overview of AI's Environmental Impact

AI technologies span across a vast landscape of use cases and models, ranging from simple regressors to large reasoning models. It is, as such, natural that their impact has a vast range across use cases. [68] has shown that the consumption of AI use cases ranges from 3.46×10^{-8} kWh for a tabular model to 9.58×10^{-2} kWh for a large agentic model. This gap in consumption in inference leads to vast differences in impacts, and where in the lifecycle they happen, with larger models having a much higher impact at inference time. This growth has heavily impacted data centers, US data centers produced 105 million tons CO_2eq in the past year with a carbon intensity 48% higher than the national average [90]. Their impact is not limited to CO_2 and key environmental indicators include:

- Green House Gas (GHG) emissions. The energy used to run the servers and build the server components emits GHG. Those GHG emissions are measured as an equivalent mass of CO_2: for any gas, it is the equivalent mass of CO_2 that has the same global warming potential as the mass of that gas, it is measured in $kgCO_2eq$.
- Abiotic Resources Consumption. These are the metallic and mineral resources needed to manufacture all the hardware to run AI and store the data. The depletion of resources is measured as the equivalent mass of antimony.
- Water consumption. Water is mainly consumed during the hardware manufacturing process and during server runs to cool them.

3.2 Generative AI's Ecological Impact

Generative AI exacerbates the environmental footprint of digital technologies across all life-cycle stages (manufacturing, distribution, use, and disposal). It consumes more electricity and resources than traditional AI tasks:

- Energy Consumption: AI's energy footprint depends on factors like data center location, energy mix, model complexity, and training duration. The growing demand for AI also stresses power grid infrastructure, with transformer supply struggling to meet demand.[66] US data centers already consume more than 4% of US demand [90], a figure expected to rise sharply.
- Water Consumption: AI systems consume water for cooling servers and generating electricity. For example, 20-50 ChatGPT requests use 500 ml of water. By 2027, AI-related water demand could reach 6.6 billion cubic meters annually. Water usage varies by location, with some data centers being more water-efficient than others. [182]
- Pollution and Biodiversity: Data center construction and operation contribute to habitat destruction and biodiversity loss. Concrete, a key material in DCs, is a major source of GHG emissions and requires significant amounts of sand, leading to environmental degradation. [180]
- Electronic Chips: Manufacturing chips for AI systems is resource-intensive, involving rare metals, pure water, and energy. Embedded AI, which processes data locally on devices, offers a more sustainable alternative by reducing reliance on cloud infrastructure. [243]

3.3 Rebound Effects and Potential Benefits

AI's ease of use can lead to rebound effects, where increased usage offsets environmental benefits. For example, AI can optimize fossil fuel extraction, inadvertently increasing CO_2 emissions. Additionally, the demand for new digital infrastructure and consumer attraction to innovation accelerates resource consumption and obsolescence.

However, AI also holds potential for reducing environmental footprints [149]:

- Directly: AI can monitor air quality, optimize agriculture, and simulate climate scenarios.
- Indirectly: AI improves energy efficiency in transportation, building management, and energy distribution.

4. USAGE PERCEPTIONS OF AI

Developing a frugal artificial intelligence is a matter of technical optimization but also of choice on the informed use of artificial intelligence, case by case. Artificial intelligence should only be used in cases where it is the best technique to use (compared to the others) but also because the intended use is useful, beneficial, expected by the society in which it is deployed and because the adverse effects of this use would be minimized and less than the beneficial effects. To enable this parsimonious and essential use of artificial intelligence, we propose to look at how artificial intelligence is perceived by public opinion by taking an interest in surveys that measure the awareness and use of generative artificial intelligence in France. The expectations, benefits, or fears that respondents highlight will then be discussed. Finally, we will study how the debate is articulated in French society.

4.1 The concept of artificial intelligence is well known in public opinion

The analysis is mainly based on four general quantitative studies [79; 80; 69; 78] and an open consultation with French citizens to suggest ideas for a beneficial use of artificial intelligence[5] [147].

The results of the studies may differ quite widely, but it is possible to see that there is a strong awareness of the concept of artificial intelligence and generative artificial intelligence, although this is a very technical subject. And a strong curiosity led the French people to try these tools.

On the other hand, these studies do not allow for the dissection of their understanding of artificial intelligence. It should also be noted that all studies are conducted online[6].

The following tables summarize the answers of different studies on two questions: Do you know generative artificial intelligence, and have you already used these tools?

Study	Question	Results	Comments
Viavoice for SSII (February 2024)	Question asked without any explanation	65% Yes	Institute comment: "65% of the French have already heard about generative artificial intelligence, a notoriety that remains, however little built, only 22% of the French see very well what it is"
IFOP for Talan (May 2024)	With explanation[7]	78% Yes	Institute comment: "Generative AI is gaining notoriety among the general public (78% have already heard of it this year compared to 71% in May 2023)"
Ipsos for CESI (January 2025)	Question: Do you know generative AI tools?	88% Yes	

Table 1: Synthesise of studies results on awareness

Study	Question	Results
Viavoice for SSII (February 2024)	Question: have you ever used a generative artificial intelligence solution?	17% Yes for personal purpose - 19% Yes for professional purpose
IFOP for Talan (May 2024)	Question: do you personally use generative AI tools	25% Yes
Ipsos for CESI (January 2025)	Question: do you use generative AI tools?	39% Yes
IFOP for Orange - Sociovision (2024)	Question: Have you ever asked some questions to a generative AI?	48% Yes

Table 2: Synthesis of studies results on use

4.1.1 Awareness of generative artificial intelligence

A huge awareness of generative AI, even if what hides behind this awareness cannot be analysed with those studies (see Table 1).

4.1.2 Use of generative AI tools

Awareness is not only a theoretical one, as more and more persons try these tools. However the gap between awareness and usage is still huge (see Table 2).

4.2 A growing media presence, but still below the major topics of society

The presence of the subject in the media sphere has grown strongly in recent years, however, it is necessary to relativize the place that the subject occupies. Indeed, a 10-year analysis of the place of the subject in traditional media[8] (see Figure 2) shows that while artificial intelligence is mentioned more and more often, and especially since the introduction of ChatGPT on the market, this presence remains relatively modest compared to other topics identified as concerns of the French people, such as immigration, climate change or purchasing power.

4.3 What are the usages of generative AI?

[5]This citizen consultation - What are your ideas for shaping AI to serve the public good – was conducted by Make.org for Sciences Po, AIandSociety Institute (ENS-PSL), The Future Society, CNum, as part of preparatory work for the Artificial Intelligence Action Summit, held in Paris in February 2025.

[6]The methods of collection for the Viavoice study are not specified.

[8]Analysis from database ina[63] on the keywords "artificial intelligence", "climate"; "purchasing power", and "immigration". The media analysed are: JT (Arte, France 2, France 3, M6, TF1), continuous information channels (6h-0h range of BFM TV, CNews, LCI, franceinfo, iTele), radio stations (6-10h range of Europe 1, France Culture, France Info, France Inter, RMC, RTL, Sud Radio). Occurrences are counted as the number of rounds in which the word was detected at least once by the IA. For example, if a word is said twice by the same person without being cut off by another person, that word will be counted once. To compensate for the disparity of time slots between media, the absolute values were indexed by taking the value of the immigration theme in 2015 on each type of medium as a base 100. An arithmetic average of the indices was then made.

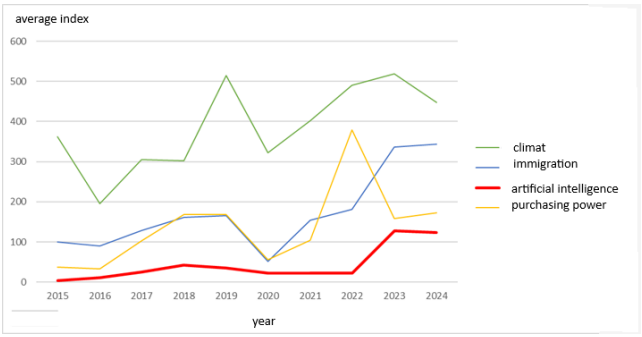

Figure 2: Topics mentioned in French media

As AI is a social definition regarding the latest technology on the market (see Section 2), studies in 2024/2025 focus on generative AI.

Beyond the awareness of the word and the concept or use, studies allow us to identify what is the social acceptability of artificial intelligence itself and its uses. The Sociovison study details the perceived usefulness of generative artificial intelligence tools. Two-thirds of the people who asked questions to AI find it useful, and the younger they are, the more urban and high-income, the more useful the use of AI is considered to be. In the professional field or for students, the use of generative AI also seems to be beneficial. With the idea of an assistant that saves time for low-added-value tasks or summarizing and synthesizing information. The Ipsos-CESI study adds translation to these most common uses. Other uses are emerging (for almost one-third of the people using AI in their trade): acquiring or compensating for a lack of skills or even making decisions.

AI is mostly seen as a human assistant, but with a significant impact on society.

4.4 Benefits and threats: a clear apprehension by respondents

There is no mention in the studies of the environmental impact of artificial intelligence, either by energy consumption, by the construction of data centers, or by the manufacture of machines. This theme is not offered to respondents. This makes it invisible. And since it is not proposed, it is not commented on, and the question does not feed into public debate. It is a general problem of the digital world whose environmental impact is not very visible.

However, the citizen consultation (Make.org) identifies 5% of proposals to raise awareness and reduce the environmental impact of artificial intelligence. The proposals are around weighing up the benefits in terms of the environmental damage caused. Artificial intelligence can also be used to monitor and thus prevent the risk of disasters or environmental degradation.

Also, the benefits and threats associated with the deployment of artificial intelligence are more related to societal impacts. The themes concerning the benefits and threats of artificial intelligence are fairly homogeneous between studies. As these mainly deal with generative artificial intelligence, they focus on this part of the technology. The quantitative surveys propose categories to people who vote on a Likert scale, according to whether they agree with this theme and its formulation. However, the open consultation on behalf of Sciences Po, by make.org, allows spontaneous themes to emerge; It should be noted that they are close to the themes assisted by quantitative studies.

The expected benefits are of several orders. First of all, we have seen above a benefit to be assisted to perform tasks with low added value in their personal (Viavoice) and professional (Viavoice, Sociovision) lives, and synthesize the information received in their professional life (Sociovision). But also, get advice or help to solve a problem as a customer (SocioVision).

Security benefits are also seen: either to obtain reliable data (the first reported benefit for respondents of the SocioVision study) or to secure navigation (by blocking malicious content). Moreover, a more specific study on the use of artificial intelligence for the French administration shows that it is mainly expected in the sectors of Defence, security and surveillance (44%,) to strengthen the fight against social and tax fraud (51%), public security, and crime prevention (45%).

The respondents of the SocioVision study, expect as a benefit to have access to reliable information; notwithstanding do they fear **not having access to this reliable information any more** (75% - this is the highest percentage among the different countries tested[9] in this study, to note that none goes below 61%, except China, to 45%). This concern is also major in the ViaVoice study for SII (83%) as the Ipsos-Cesi study (49%)[10]. The latter also identifies a risk of loss of discrimination between what is real or generated by AI (43%) and even the use of false or unreliable data. Among the risks of using AI by the administration, respondents from the Ifop/Acteurs public study point to the risk of error of these AI.

The second threat, very strongly identified, is that of the decrease in contacts between people (SocioVision), **the dehumanization of social relations** either from a general point of view (ViaVoice) or in relations with the administration (Ifop/ Acteurs Publics).

This nuanced vision of the integration of artificial intelligence in different areas of personal and professional life leads respondents to prefer a deployment framed by regulation en-

[9]Germany, Spain, Poland, United Kingdom, USA, China, Morocco, Egypt

[10]The rates are very different between the ViaVoice study where the themes of concern identified are all approved by a range between 63% and 83% of respondents (the question is: in the future, Do you fear the rise of artificial intelligence? Do you think that they do not allow you to tell the difference between true and false in terms of information?) and the Ipsos Cesi study, in which no concern concerns more than 49% of respondents (the question is: In your opinion, what are the main risks associated with the use of generative AI? *among the proposals:* The spread of false information (fake news). It is not specified how many choices respondents could make.

acted by public authorities. This is the case for 86% of the French respondents in the SocioVision study (this rate is similar in all the countries tested and ranges from 78% in Germany to 90% in China. Note that the Americans approve of the need for regulation at 80%). The team[11] analyzing the citizen consultation on Make.org explains this request: "Participants reject any form of AI solutionism and uncontrolled deployments. Participants call for robust governance frameworks, both at the local and international levels, to safeguard their rights and protect human agency. They are divided about unchecked deployments of AI systems and reject the idea of leaving key decisions to private companies".

4.5 A nuanced debate on the part of civil society, and polarized by actors in the field

Section 2 of this document shows that artificial intelligence remains a vague and ambiguous concept. Using this notion to feed the public debate erases technical expertise to put questions on the overall functioning of society. This has two implications for public debate.

First of all, it facilitates **the inclusion of the citizen in the debate**. The analysis of citizen consultation in France for the Action for AI summit, early 2025, allowed a first debate (approval/ rejection of proposals). The results show that it is possible to have a fairly measured debate. For example, proposals under the "Stop the AI" theme, which is a clear-cut position, are controversial and received approval and rejection votes in roughly equal proportions[12].

The second consequence is the counterpart of this conflation. Indeed, the actors of AI and especially the entrepreneurs of the Silicon Valley rely on the credibility that their knowledge of the subject gives them to take very global positions on the future, such as the ones quoted by Heaven [99]:

- Marc Andreessen: "This has the potential to make life much better [...] I think it's honestly a layup.
- Altman: "I hate to sound like a utopic tech bro here, but the increase in quality of life that AI can deliver is extraordinary."
- Pichai: "AI is the most profound technology that humanity is working on. More profound than fire.".

[11] Constance de Leusse, AI & Sociéty Institute (ENS-PSL) and SciencesPo Tech & Global Affairs Innovation Hub; Nicolas Moës, The Future Society; Axel Dauchez, Make.org; Jean Cattan, National Digital Council; Caroline Jeanmaire, The Future Society; Tereza Zoumpalova, The Future Society; Alexis Prokopiev, Make.org;Marthe Nagels, Make.org; Victor Laymand, Make.org; Pierre Noro, SciencesPo Tech & Global Affairs Innovation Hub; Mai Lynn Miller Nguyen, The Future Society; Niki Iliadis, The Future Society; Jules Kuhn, Make.org

[12] This consultation is not representative of the opinion of the French population; it does not involve interviewees on each proposal or a representative sample, but people who have voluntarily joined the consultation, Draft suggestions and, on the other hand, evaluate the agreement or rejection of other suggestions made. The proposals judged are not exhaustive: everyone chooses those on which he or she decides. Over 11,000 people participated.

Making artificial intelligence a total tool highlights potential apocalyptic risks for humanity. And focus the reflection on these existential risks instead of facilitating a calm debate that would help to understand what companies want to build as a future with this technology, causes opposition between "**accelerationists**" (to accelerate deployment, seek it with the conviction that the benefits will always be greater than the disadvantages) and the "**catastrophists**" (demanding a halt (or a moratorium) in the face of incalculable and existential risks for humanity).

Thus, this opposition prevents us from truly thinking about what AI is doing to societies. Charlie Wazel is a journalist who investigated how the actors of the Silicon Valley (here around OpenAI) present their work on artificial intelligence. His article, published in July 2024 in The Atlantic, is entitled "**AI has become a technology of faith**". He writes: "In this framework, the AI people become something like evangelists for a technology rooted in faith: Judge us not by what you see, but by what we imagine [218]" .

This prevents us from thinking about the concrete problems that are already there, and that the hope of the future cannot be sufficient to sweep away [94]. This also allows established actors to thwart regulatory projects: "Thus, the big tech players are readily in favour of a desire for regulation that would focus on the apocalyptic risks for humanity, coming from the innovations of "frontier" and less on their own model [30]".

4.6 A polarization of the debate that is detrimental to thinking

Many risks are well identified by citizens (see the perception of risks in the various studies, described above), but some are invisible because they are not proposed to respondents and therefore not taken into account in the analyses. These include the environmental consequences of these technologies (see chapter 3) or the work of people who feed artificial intelligence or correct it [159].

The citizen consultation organized for the AI Action Summit allowed respondents to contribute to the debate. On the other hand, in the context described of a vague notion, totalizing or even considered as magical that oscillates between vital necessity and apocalypse, the use of surveys to measure public perception is part of a process to work on social acceptability and not on democratic reflection on the subject of artificial intelligence.

The presentations made in the studies or their analyses show that artificial intelligence is obvious, which prevents us from thinking about it. This is what Julien Falgas and Pascal Robert describe in The Conversation, taking up their concept of "unthought of the digital [75]". The studies that have been taken up at the beginning of this chapter are part of this vision of an obvious, the progress that constitutes artificial intelligence, and on the necessity that all "start". The words used in the texts are directed to this objective.

- In the SocioVision study, the issue described that motivates the questions around artificial intelligence is: "the issue: putting generative AI at the service of progress for all."

- Similarly, the IFOP-Talan study comments on the results as follows[13]:
  - Generative AI is **gaining** notoriety,
  - Their use remains minority but is **making progress**.
  - Generative AI seems to be more **democratized** in working life.

- ViaVoice, for SII, comments on the results as follows: ViaVoice for SII: " Artificial intelligence solutions appreciated by **insiders**" and "due to this still **poorly knowledge**, the rise of artificial intelligences worries the majority of French people"

- Finally, EY draws up recommendations for public sector actors, based on the study conducted by the Ifop) with the following assumption: **"If there is no longer any need to demonstrate the value of adopting AI in the public sector, it is important to understand what are the key success factors to have it adopted"**. The recommendations detail ways to build public confidence. The first is **acculturation**, the next two are more technical, and finally, the last targets the necessary regulation.

These various quotations are intended to show that the vocabulary used by those who animate the debate is already marked by the solutions they wish to push. And as the critic Guy Marcus, a champion of generative models but promoter of more diverse artificial intelligence: "Neural network people have this hammer, and now everything is a nail" says Marcus[99].

This section aims to understand the perception of artificial intelligence in public opinion through quantitative studies (surveys) and propose a critical reading. Indeed, surveying is not participation or debate. Then, the experts reappropriate the opinions expressed to propose policies that allow, as we have just seen, finding the best ways to deploy artificial intelligence without necessarily questioning society's expectations and taking the risk of not analysing the consequences of this deployment globally (forgetting precarious workers and the environment, for example). But working with the public and civil society to shape the intended use of artificial intelligence, rather than making it a matter for experts, could only be beneficial in taking seriously the skills of people who will be affected by this technology. Indeed, as suggested by the make.org consultation team: "The public opinion demonstrates a sophisticated understanding of AI. Participants are numerous and demonstrate nuanced and diverse opinions of AI's potential and risks. Despite the technical nature of the matter, the level of awareness validates the importance of involving the public and civil society in the governance of AI.

[13]highlights are from the author

5. ECONOMIC FORECASTS: AI AND FRUGAL AI

5.1 Preamble - Context

AI has become a central pillar of economic transformation. However, the debate between energy-intensive AI models, and more efficient FAI (ie Frugal Artificial Intelligence) approaches continues to shape investment strategies, adoption trends, and operational costs. Let's examine the economic outlook for both AI paradigms in the next five years, analyzing supply and demand dynamics, labour market implications, and the way time-to-market constraints contribute to bolster the not-always relevant all-LLM trend.

5.2 The Supply Side

The implementation of FAI depends largely on the economic conditions affecting AI services. This includes factors such as industry investments, profitability expectations, market consolidation, pricing strategies, and resource constraints.

5.2.1 Industry Investment

On the ground of profitability, the AI industry has witnessed significant capital inflows, yet many leading AI firms are operating at a significant loss to gain market share. OpenAI epitomizes this situation, reportedly spending near $700,000 per day to run ChatGPT [73], at least over a certain period. Profitability horizons remain thereby uncertain due to high operational costs. The recent arrival of allegedly far more efficient challengers such as DeepSeek [134] brings in this landscape its own share of extra uncertainty. This exceptionally competitive environment leads actors to deploy unusual efforts of persuasion to depict AI-based services as an inexorable necessity calling for fast adoption.

Furthermore, given the enormous cost of developing and running LLMs, market consolidation is expected in the coming years [67]. Larger tech firms are acquiring AI startups to integrate new technologies quickly. This trend may be logically expected to continue in the coming years as smaller firms struggle to compete with industry giants. But some of the latter may still have to prove they don't stand on feet of clay, when cheaper competitors burst in the place.

5.2.2 Resource Constraints

Eventually, the constraint on resources can become a pivotal issue for the supply side. AI models require vast computational resources, particularly GPUs and energy. The demand for AI data center capacity is expected to triple by 2030 [89]. This could create bottlenecks that impact pricing and access to AI services, potentially increasing demand for more energy-efficient alternatives (depending on the case, cheaper SLMs or -wherever applicable- pure FAI with no generative capacities).

5.3 The Demand Side

The adoption of AI services varies among professionals and the general public. While demand is growing, key barriers include cost concerns, model reliability, and integration challenges.

5.3.1 Professional Adoption Trends

Regarding trends of the professional segment, enterprise adoption of AI is accelerating, with surveys indicating that 65% of companies now use generative AI regularly [52]. However, this adoption copes with two impedimenta. First, LLMs, because of explainability and/or latency issues, simply cannot suit every industrial or educational need, even where they are theoretically relevant. Second, the cost of running LLMs without enough selectivity may sometimes turn into OPEX explosions and encourage businesses to seek more efficient alternatives. The way arbitration may take place will be discussed in sections below.

5.3.2 Consumer Adoption Trends

For the consumer side, AI applications have grown rapidly, with ChatGPT reaching 100 million users within two months of launch [105]. Despite this, cost pressures and the introduction of subscription fees may affect long-term consumer adoption, especially in case of an economic downturn induced by both Chinese [132] and American [192] contexts.

5.3.3 Cost-Effectiveness and Reliability

LLMs provide unparalleled flexibility but at a high cost per inference [107]. FAI, when applicable (namely, when the output does not call for a generative approach implying a "decoding" part), offers an alternative with not only lower operational expenses but sometimes greater accuracy and shorter latency, making it attractive at different regards, and especially, but not only, for enterprises with budget constraints.

5.4 AI vs Human

The economic impact of AI on the workforce is a crucial consideration. While AI enhances productivity, concerns over job displacement persist.

5.4.1 Workforce Displacement

On the one hand, AI automation is projected to replace approximately 300 million full-time jobs worldwide [120] - not to mention the prominent example of the Qingdao Port, already close to be an unmanned site fully automated by a mix of AI technologies and 5G networks, achieving continuous records of performance [238]. On the other hand, new roles in AI development, oversight, and management are expected to emerge. The medium-term horizon of this Schumpeterian scheme is at this stage highly unpredictable, given its political "unthought" and the plausible limitations coming from energy and natural resources.

5.4.2 Human Competitive Advantages

Despite AI advancements, human expertise remains critical in areas requiring emotional intelligence, strategic planning, and interpersonal communication. Yet, creativity should no longer be perceived as a human turf but rather as a battlefield with local victories [124], perennial or not. That said, not ignoring the emerging "reasoning" capacities of cutting-edge LLMs, human induction is probably not immediately threatened on the short term, especially when it applies to the perception and the understanding of reality. Galileo stated the law of uniform motion in purely counterfactual reasoning, without any statistical arsenal, and moreover never having been able to produce the experimental vacuum. A constrained world may sharply foster these cognitive abilities.

The IT sector stresses a specific set of questions. Will software development as-we-know-it steadily disappear, as foretells Nvidia CEO [51]? In an "infinite world", the question has its share of legitimacy, except probably for technological or military processes constituting an existential issue. By the end of the decade, will data science skills experience similar shifts, with 80% of machine learning tasks likely to be automated [48]? Likewise, such assumptions - partly relying on the progress of "AutoML" services that epitomize an energy-intensive philosophy, will have to be updated in the light of energy reality and the subsequent trade-offs.

5.5 Conception vs. Run Costs

A key distinction in IT economics lies in the difference between the conception phase and the operational (run) phase.

5.5.1 Generative AI Accelerating Conception

Gen-AI significantly reduces the time required for ideation and prototyping across industries. For example, product designers can rapidly iterate concepts using AI-generated mock-ups. In numerous situations, Gen-AI can also deliver a dramatically easy implementation of functions-as-a-service (FaaS). Indeed, if N-tier architectures enjoyed a great comfort of conception with interface definition frameworks during the last decade (e.g. OpenAPI), micro-services, per se, can be now easily implemented with Gen-AI integrated solutions [122] or through basic software craftsmanship (e.g., prompting for structured JSON objects).

5.5.2 Run-Time Costs

However, magic has its drawbacks. Operating LLMs incurs significant computational and energy costs. Studies show that for specific non-generative tasks (e.g. natural language classification) where FAI or vanilla algorithmics can pretend to compete with, and sometimes outperform, LLMs, the latter can have an energy consumption significantly higher [152] (with, thereby, similarly higher carbon emissions).

Hopefully, the combination of optimization techniques like cascades, approximation, and prompt adaptation can theoretically save a significant percentage of energy in eligible situations [144]. Are organizations, though, always in the practical conditions to spend resources on such efficiency improvements? The answer is not self-evident as long as we live in a world of cheap and abundant energy, and where the relationship to time is a predominant determinant of economic competition.

5.5.3 Naive Time-To-Market (TTM) pattern

The development cycle of a software feature is often TTM-driven due to the competition for the early conquest of the largest market share. When (and only when) the foreseen functionality is deemed eligible to frugal algorithmics, comes most often a dilemma. Develop an accurate, reliable, tailor-made FAI-based solution (calling for labelled data, model training, high skills and a longer conception phase)? Or implement, faster and probably with a reduced development team, a Gen-AI-based approach? The two scenarios are rep-

resented hereafter with their respective timeline (see Figure 3). Let's underline that the schemes are purely didactic, so as to depict the cost distribution likely to happen in each situation. First, they do not reflect real figures. Second, they do not embark specific conception approaches like fine-tuning or similar techniques.

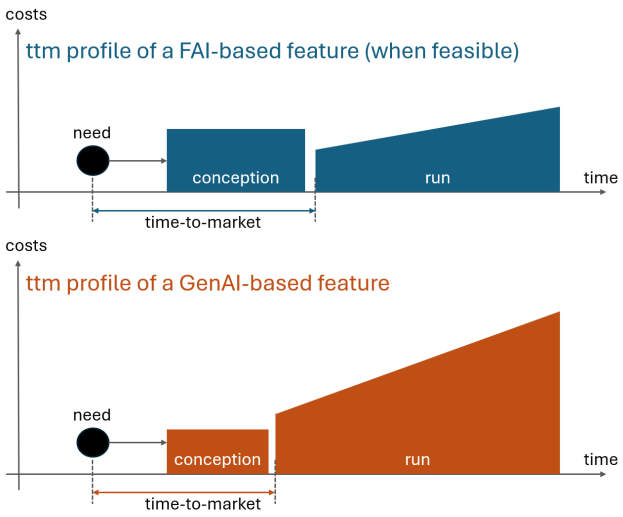

Figure 3: The timeline dilemma: launching faster or building smarter

In short, with irrelevant Gen-AI usage involved in runtime platforms, costs can rise faster than expected, putting product pricing at risk. This may be yet justified by a strategic effort to secure early adoption by a market segment.
In those cases, though, appears a challenge of project governance: to apply a proactive and frequent assessment of the relevance of Gen-AI usages at runtime. A modular software architecture (e.g. micro-services) with well-documented interfaces is the cornerstone of such continuous improvement efforts.

5.6 Summary

In the next five years, economic factors will drive AI adoption choices. While LLMs continue to enable groundbreaking innovation, their high operational costs may push organizations toward FAI solutions, especially in the present geopolitical turmoil, where several clues indicate the closer proximity of a world governed by finitude, especially at the turn of the next decade [203]. A balanced approach, leveraging the strengths of both paradigms, is likely to define the future of AI deployment.

6. PLANET BOUNDARIES - ON AI DEVELOPMENT AND ENERGY RESOURCES

6.1 Growth in the use of AI

Most observers estimate [86] that growth in usage and associated sales will follow an exponential curve, at least by 2030. This growth is underpinned by a particularly rapid rate of adoption of AI compared with that observed for other, equally recent technologies, in which it is indeed generative AI that is driving this growth in AI usage [155].

This growth requires the associated material equipment in the form of servers providing the necessary memory, power, and computing speed [178], the manufacture of which implies the availability of natural resources (water, metals, etc.), and the operation of which implies the availability of the required electricity.

6.2 Electricity resources required to operate the AI, needed to sustain AI growth

6.2.1 Evaluation to 2030

The growth in electricity required to operate the corresponding data centers will follow a more moderate curve than that of AI usage, thanks to energy and architecture gains [112]. However, these (linear) gains will not compensate for the growth in electricity needed to keep pace with demand.

The United States [24] has estimated a projection of data center consumption between 2024 and 2028, according to two scenarios (high and low), which include, on the one hand, the growth in storage and computing power, and on the other hand, these energy gains.

Between 2010 and 2022, global electricity production grew by 50%. Between 2022 and 2040, it should grow by 100%, i.e. double, and then increase by a further 25% between 2040 and 2050 [176], corresponding to linear growth from 2010 to 2050.

An admittedly simple model (approximation of the growth in energy requirements by an exponential curve, see Figure 4) based on the data for 2024 and 2028 mentioned above for the USA, scaled up to the global level (the USA consumed 17.3% of the world's electricity in 2023 [113]), of electricity consumption by data centers, using an average scenario built as the average of the two scenarios (LC and HC), leads to the Table 3.

2020	2021	2022	2023	2024	2025
0.863	1.056	1.2813	1.6	1.979	2.455
2026	2027	2028	2029	2030	
3.052	3.805	4.754	5.951	7.464	

Table 3: % electricity production used by data centers, Medium case

AI is not specifically discerned in this assessment, however, it has been noted that the preponderant (exponential) part of this growth is linked to the use of generative AI. According to this modelling estimate, by 2030 7.5% of the world's electricity production would be consumed by data centers.

6.2.2 Evaluation beyond 2030

The use of data for projections beyond 2030 is risky, due to the scarcity of data and the high degree of uncertainty surrounding the evolution of other resources likely to support growth (metals in particular), as well as the growth in computing requirements linked to AI. Unsurprisingly, however, it would reveal a divergence between (linear) growth in electricity production and (exponential) growth in data center consumption (See Figure 5).

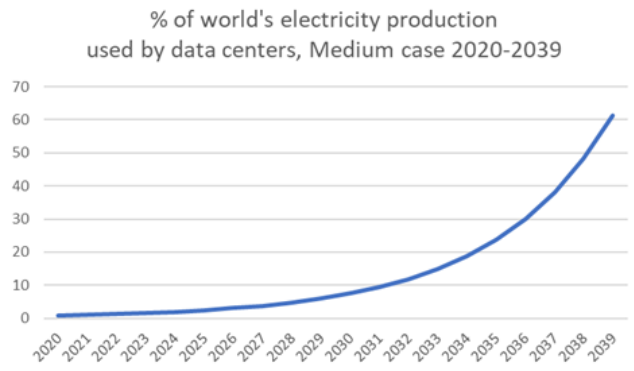

Figure 4: % World's electricity production used by data centers

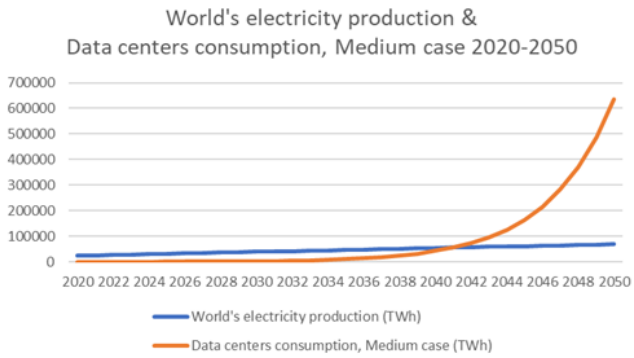

Figure 5: World's electricity production & data centers consumption

In particular, all the electricity generated in the world would be consumed for data center needs as early as 2041.

6.3 Analysis

6.3.1 Conflicts over electricity use

Electricity, a limiting factor for AI growth - The growth of AI, through the surplus electricity it requires, will be confronted with its need for energy as a limiting factor in this growth. At the same time, it will intensify conflicts over the use of the electricity produced, which, barring a technological breakthrough (controlled nuclear fusion in particular, under research since the 1960s), is unlikely to be able to sustain this development. This raises the question of arbitration between different economic players regarding the availability of electrical energy resources.

The position of economic players and the search for new sources of electricity generation - The conditions for maintaining economic activity will then be, in addition to the control of one's own production processes, that of access to electrical energy. This analysis explains why some major electricity consumers are already seeking to secure their electricity supplies, in particular by:

- privatizing production centers (e.g. units in conventional nuclear power plants [202]);
- deploying their own means of production (solarization) [43] ;
- investing in, or forming partnerships with, innovative power generation facilities such as nuclear Small Modular Reactors (SMRs), which can be adapted to keep pace with the growth of a data center [198].

From this observation, we can also see that the economic activities that will best be able to maintain themselves over the long term will be those that have secured their electricity supplies, either through direct control of their own electricity production facilities, or through a certain financial capacity by going to the electricity financial markets.

6.3.2 Focus on France

Between 2035 and 2045, about half of France's current nuclear power generation capacity will no longer be available. Nuclear power plants, built in comparable years under the auspices of the Messmer Plan, are located on water-stressed rivers, and most of them will not be able to be maintained beyond 50 years [81].

7. USE THE RIGHT AI FOR THE RIGHT NEED AT THE RIGHT TIME

7.1 Preamble - Life cycle of an AI system

The life cycle of an AI system is similar to the old one named "life cycle of data mining project" [135]. In this section we are interested in Lifecycle Assessment (LCA) [126] which is a systematic approach to evaluate the environmental impacts of a product or system throughout its entire life cycle[14]. As for data mining, the AI lifecycle encompasses the complete process of developing and deploying artificial intelligence systems. It starts with data collection and moves through stages such as data preprocessing, model training, evaluation, deployment, and ongoing monitoring and maintenance. For more details on standardization see Section 10.

Due to the Life cycle of an AI system, here is a list of the costs that prevent the AI from being frugal[15] (a non-exhaustive list): (i) Development Costs (ii) Data Costs (iii) Infrastructure Costs (iv) Training Costs or retraining cost (v) Inference cost (vi) Maintenance Costs (vii) Compliance Costs (viii) Deployment Costs (iX) Support Costs, etc. These costs can accumulate and impact the overall frugality of an AI system, and the reader may find more details in recent publications as for example: [225]. The cost to pay is the addition of these costs (and some of them have to be paid at every use of a given model as for example the inference cost). Contrary to some publications, the cost to pay is not

[14]We do not study AI-enhanced LCA models which try to improve the precision and depth of environmental impact assessments [21].

[15]We do not define frugality here, see section 2. But we can think in this section that total costs can have a minimum value given a task to be solved and an ROI to be achieved. In this sense, the idea is to try to get as close as possible to this value.

only the three steps: training, deployment, and production. We encourage considering the sum of all these costs and not only part of them (for example fine-tuning[16] of the existing model only reduces one of the costs (the training cost)). Even when only the model has to be updated, potentially updating the model is an investment decision which, as in the financial markets, should only be taken if a certain return on investment is expected [245] and frugality should be taken into account.

Another point in this period is the use of large models (Generative AI, large deep neural networks, etc.). It could be interesting to keep in mind that "old models"[17] particularly on Tabular data or Time series remains quite interesting in terms of performances (see the example below in section 7.3).

The list of tasks that could be performed with AI is very large (classification, regression, etc). Many of them are currently not frugally solved by large models. Indeed, one of the key points in frugality is finding the right inflection point between performance and frugality (all the cost to pay), which is the focus of the next subsection.

7.2 Finding the right inflection point

Finding the right inflection point between performance and frugality indicators in AI models is critical to maximizing efficiency, accessibility, and ethical considerations, while still achieving satisfactory levels of performance. Balancing these factors can lead to more sustainable and impactful AI solutions. There are many arguments in favour of finding the right tipping point[18], but here are a few of the more obvious ones:

- Resource efficiency:
  - Cost reduction: Energy-efficient models require less computing power and memory, resulting in lower operating costs.
  - Environmental impact: Reducing resource consumption can reduce the carbon footprint associated with training and deploying AI models.
- Scalability:
  - Broader accessibility: More efficient models can be deployed in resource-constrained environments, making AI accessible to a wider audience.
  - Faster deployment: More efficient models can be trained and deployed faster, allowing rapid iteration and adaptation.
- Optimized Performance:
  - Diminishing returns: At a certain point, increasing model complexity yields minimal performance gains. Identifying the tipping point helps avoid unnecessary complexity.
  - Robustness: Simpler models can sometimes generalize better to unseen data, reducing the risk of overfitting.
- User Experience:
  - Latency reduction: Frugal models often result in faster inference times, improving the user experience in real-time applications.
  - Ease of integration: Less complex models can be more easily integrated into existing systems and workflows.
- Ethical Considerations
  - Fairness and transparency: Simpler models can be more interpretable, making it easier to understand the decisions made by AI systems and promoting fairness.
  - Bias mitigation: Frugal models can reduce the risk of embedding biases that can result from overly complex architectures.
- Innovation and experimentation: Encouraging creativity: A focus on frugality can inspire innovative approaches to problem solving, leading to novel solutions that may not rely on heavy computational resources.
- This list is not exhaustive, of course, and we can add costs that are sometimes 'hidden', such as increasing the skills of teams, integrating an additional data scientist into the project team, etc.).

One way to find this trade-off is to use benchmarking [61], which plays a crucial role in the development of frugal AI by improving efficiency and adaptability. The results of benchmarking AI methods help to develop more frugal AI in several ways. Firstly, it is possible to identify efficient methods, since benchmarks enable comparing the performance of different AI methods, highlighting those that offer the best value for money in terms of the resources used. Secondly, it is possible to optimize resources: through analysis of the results, researchers (i.e. users) can identify algorithms that require less data or computing power, thus favouring lighter solutions. They also provide a consistent framework to evaluate AI models, ensuring comparability across different approaches (standardization). They help identify the most efficient algorithms for specific tasks, guiding resource allocation (performance metrics). They encourage sharing of best practices and datasets, fostering innovation in frugal AI solutions (community Collaboration).

Note: The aim of benchmark results is not to systematically compare solutions (by repeating a lot of experiments), but to build up a set of skills that will enable an appropriate selection to be made. The question is therefore "how can companies that do not have data scientists build up this knowledge" (or companies that have qualified data scientists but who are overloaded with work and therefore cannot respond to all requests, etc.).

7.3 Illustration on sentiment analysis

As far as we know, there is no universal method for finding the right tipping point. Modestly, however, we can mention one that makes sense at the start of a data science project: (i) define the performance criterion for the project; (ii) define the value of this criterion (perhaps in the form of a return

[16] See Section 14.2 for a definition of fine-tuning

[17] We mean by 'no large models' as for example Linear Regression, K-nearest neighbours, Random Forest [33], Catboost [177], Khiops [29], etc. or even signal processing for time series as, for example, exponential smoothing, Arima, etc. [31]

[18] This can also be seen in terms of simplification gains.

on investment (ROI)); (iii) use a rule, an AI, etc., that is simple at the start and then, if the value of the criterion is not reached, make the AI more complex; (iv) stop as soon as the value of the criterion is reached or when the sum of the costs becomes too great (or the return on investment cannot be achieved or the cost of achieving it will be too high).

This is illustrated in Figure 6: In the purple case, if the return on investment in terms of performance is achieved with P1, there is no reason to make the AI more complex and pay additional costs. In the green case, the same performance can be achieved for two different costs. It is therefore very interesting to start by using an AI producing cost C1 and then stop. The worst case is where using an AI produces a higher overall cost with poorer performance (not illustrated in the figure).

This last scenario is well presented in [153]. In this report a classification task is designed on text (sentiment analysis) using a Support Vector Machine (SVM) [57] or three Large Language Model (LLM)[19]. For this given classification task we may observe that the biggest LLM energy consumptions for inference are they are several orders of magnitude higher than a standard SVM for a comparable (or lower) accuracy.

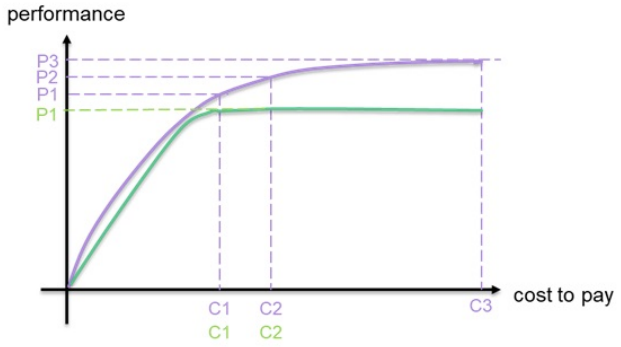

Figure 6: Illustration of different tradeoffs between performances and costs

8. ASSESSMENT OF ENVIRONMENTAL FOOTPRINT OF AI

8.1 Life Cycle Assessment

To reduce the environmental impacts of AI, those impacts need to be identified and measured [18]. Methods relying on Life Cycle Assessment (LCA) (see Figure 7), as defined by ISO 14040 and 14044 standards, have been proposed in [146]. Impacts exist throughout the life cycle.
The variables that influence the environmental footprint of AI, discussed in detail in Section 3, must be kept as low as possible throughout the AI life cycle. This section focuses mainly on machine learning aspects rather than symbolic AI (see, e.g. [83] for a symbolic AI definition and its relation to machine learning), except for some tools given in the latter case.

[19](BERT fine-tuned on the problem to solve, Llamma and BERT prompted to solve the problem)

The life cycle of machine learning AI systems [64] consists mainly of:

- Collecting, storing, and preprocessing data,
- Training and assessing models with the previously collected data,
- Running the best models in applications.

It should be noted that these steps are not fully sequential and may be interleaved, e.g., new data may be collected while running the system to train new models.

8.2 Energy consumption: challenges

Today, there are three major research challenges linked with energy consumption in AI:

- Defining unified measures for energy consumption of various algorithms.
- Evolving measures sideways with the emergence of new AI methods.
- Determining correlations between measurable variables (e.g., energy consumption, carbon footprint, greenhouse gas) and major political and industrial efforts.

To reduce the energy consumption of AI training and inference, it is critical to develop a common measurement framework that includes a complete system, as well as a per-component energy evaluation. The objective is to identify components prone to optimization and compare different algorithms.
Today, there is no unified tool that evaluates these steps for all use cases, usages, and data types. Recent research efforts provide training and inference evaluations of ML methods, see [187], [209] and references within.

8.3 Energy Consumption Measurements

To evaluate the energy consumption of machine learning functions and/or hardware components, one needs to define the software and hardware use case characteristics and appropriate measures associated with them. There are three categories of measurements:

- External power meter (EPM) measurements of hardware components.
- Energy profiling of physical components and/or algorithms (e.g., estimation of energy consumption based on calculus-related hardware or software variables).
- Measurements of built-in components or sensors of specific manufacturer solutions (e.g., CPU, GPU, or several hardware components).

The EPM is a baseline method for evaluating energy consumption. It is used to evaluate virtual [119] or physical systems (from integrated circuits [27] on top of specialized sensors, measurements of systems [185] by wall outlets, towards clouds [6] or large-scale data centers [170]). However, all three measurement categories have their drawbacks. For example, EPM suffers from an inability to provide the fine-grained energy assessment of methods and tools, and is costly at scale [8].

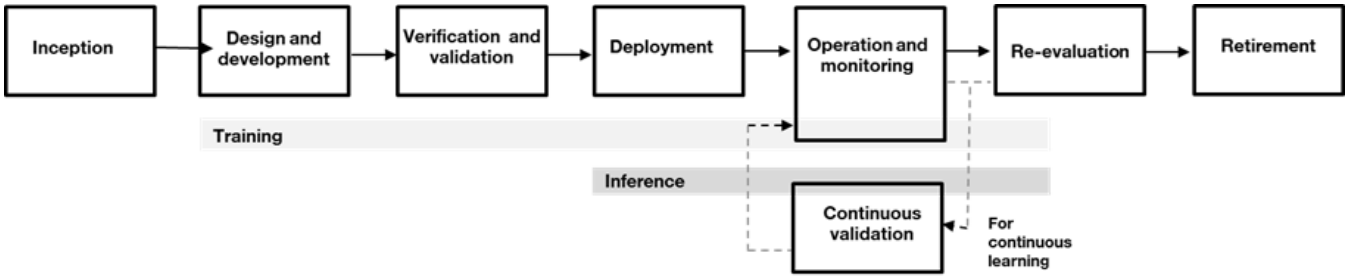

Figure 7: AI System life cycle

8.4 Greenhouse Gas Emissions Measurement

For the moment, regarding AI, Orange's internal studies have focused on Greenhouse Gas Emissions (GHG). Other variable impacts will be evaluated in the future with the same methodology when data becomes available.

8.4.1 Source of GHG Emissions

Following [146], several sources of emissions can be identified:

- Embodied emissions: the emissions associated to the production of hardware for training/inference and data storage.
- Power Consumption: the emissions due to power consumption. Electric energy consumption is used to estimate greenhouse gas emissions by Eq.(1):

$$GHG_e = C_i \times E, \tag{1}$$

where E is the consumed energy in kWh, C_i is the carbon intensity of electricity production in $kgCO_2eq/kWh$ and GHG_e are the GHG emissions in $kgCO_2eq$. C_i is highly dependent on the energetic mix. Measuring the consumption of an AI model is, however tricky as they are executed in large computing clusters. As such, it requires additional hypotheses depending on the tool used to make the initial measurement, for example:

  - If an EPM is used at node level, at least PUE (Power Usage Effectiveness: the ratio between the energy consumed by the whole datacenter and the energy consumed by computing equipment [148]) needs to be taken into account to get an approximation of the node in the datacenter, then another approximation is needed to narrow it to the model code.
  - If a code tracker such as Code Carbon [59] is used, then both the idle consumption of the infrastructure, which is the energy consumed by computing nodes when no specific computation is running on (the energy correspond to the operating system run), and the PUE should be factored in to reflect both infrastructure inefficiencies and potential under-use of computing nodes.
  - If GPU consumption alone was estimated (either through GPU-Hours, or FLOPS calculators such as LLMCarbon), then per [146] it only represents dynamic power consumption and an extra effort must be done to estimate the idle consumption and the infrastructure (network, storage, cooling, building, etc.) consumption in order to have a better estimate of the model consumption.

8.4.2 Tools

Different software tools are available to measure or estimate GHG emissions, mainly direct emissions due to power consumption during training and inference. These tools provide power consumption and convert it to GHG emissions as in Equation (1) using estimates of the carbon intensity. Power consumption measurements with software tools are not straightforward, and differences in power as measured by physical and software tools can occur, see [115].
These software tools may be generic for broad software development, or specified for a given programming language or machine learning approaches, such as deep learning or large language models (LLMs).
Here are examples of such tools:

- Code Carbon: Code Carbon is a Python library that reports CPU, GPU, and RAM consumption [59]. For CPU, on Linux, it relies on Intel and AMD Processors on Running Average Power Limit (RAPL). In Intel architectures, measurements are retrieved from registers storing physical power measures, while in AMD, they are estimates from a set of events from the core processor, IOs [115]. For GPU, only NVIDIA boards are handled, relying on NVIDIA Management (NVML) library. For RAM, a simple rule of thumb is used: 3W are accounted for per 8GB.
- ML CO_2 Impact: Machine Learning CO_2 Impact provides estimates of GHG emissions resulting from the power consumption of specific hardwares (GPUs and CPUs), using their Thermal Design Power (TDP), which gives an upper bound on the power consumption, and the duration of usage. It also takes into account the cloud provider and location of the cloud to estimate the carbon intensity of the electricity, assuming that the cloud energy supplier belongs to the same location as the cloud) [129].
- ecologits: Ecologits provides estimates of electricity consumption, GHG emissions, abiotic resources depletion, and primary energy consumption for LLMs inference. Electricity consumption is estimated for a given model and a given number of tokens. It takes into account an estimated number of GPUs needed to perform inference. It is assumed that the computing node is an AWS cloud instance with 8 NVIDIA A100 with

80GB of memory GPUs. The electricity consumption also takes into account the idle power consumption by applying a PUE of 1.2. GHG emission estimates account for both energy consumption and embodied emissions.

All these tools, even those that perform measurements while running training or inference, rely on estimations, particularly on electricity and carbon intensity. The latter two are highly dependent on the electricity provider, the time of the day, of the year, and on estimates of the carbon footprint of hardware and a hardware life expectancy. However, these tools are useful for providing an order of magnitude. If the same tool is used in an appropriate condition, it can be used to compare several hardware setups, machine learning models, and algorithms, and to assess the improvements that are implemented to decrease the carbon footprint.

In addition to those tools, cloud providers monitor the carbon footprint of the whole service of embedding AI components. Those measures are also relevant for assessing the carbon footprint of a full service, but do not provide the specific impact of AI components.

There are a variety of tools, measures, and procedures. The appropriate one must be chosen, depending on whether one wants to compute the impacts of the complete system or to deep dive into a specific component to decrease its impact. In the latter case, care must be taken to ensure that decreasing its impact does not increase the impact of another component.

9. ACCULTURATION

There is a considerable amount of work to be completed to progress beyond the initial group of individuals who are aware of and comprehend the subject. Since the release of generative artificial intelligence tools such as ChatGPT, a significant proportion of the population has become accustomed to using these tools, unfortunately, without being aware of their environmental impact. It is important that "how to design frugal AI, how to be aware of AI costs" is brought to the attention of the public, albeit with the understanding that this will require a significant investment of effort to educate and popularise it.
Acculturation to environmental impacts should be central to the implementation of Frugal AI principles, aiming to raise awareness and provide actionable tools for all stakeholders (citizens, employees, students, decision-makers, politicians, etc.).
Best practices in eco-design for AI should be integrated into existing development processes within organizations to enhance effectiveness.
The success factors for transforming organizations towards sustainability are numerous. However, it is often easier to align implementation with co-benefits such as cost reduction, stakeholder engagement, and highlighting positive impacts on the economy, environment, and society.

Here are the main Best Practices recommendations for going toward a frugal AI (see the standardization afnor for frugal AI) :

- **Challenge the necessity** and identify potential negative environmental impacts (both direct and indirect) in advance. To involve decision-makers in taking account of the challenges of sustainability and AI, (The Climate Change AI) association is catalysing impactful work at the intersection of climate change and machine learning, with a dedicated section for decision-makers.
- **Define an appropriate and frugal solution**, prioritizing traditional AI over generative AI. Select the model with the least impact that meets the needs in all cases. (The AI energy score), a joint initiative between Hugging Face and Salesforce, is a dashboard that identifies the model that consumes the least energy to perform a task.
- **Measure** environmental emissions throughout the project's entire lifecycle and share the results. To be at the cutting edge of these issues, you should follow the work of PhD Sasha Luccioni, or look at the progress of the initiative launched during the AI action summit for a global observatory on AI and energy (link...).
- **Propose continuous improvements**, such as limiting functionalities to essential needs, optimizing models, and reducing data used for (re)training.
- **Consider circularity**: reuse materials and avoid new purchases. It is noted that 45% of environmental impacts are found in data centers (Numerique quel impact environmental en-2025).
- **For GenAI solution, optimize inferences** and train users on prompts (fewer prompts lead to lower carbon emissions). There are comparators such as compare.ia, which makes users aware of the art of prompting and developing their critical faculties concerning the results obtained and energy costs.

To go further, it is recommended that these eco-design principles be combined with the principles of ethics and responsibility in order to promote a systemic view of impacts. Here is a reference that tends towards this approach, led by the French Institute of Digital Responsibility.

10. STANDARDIZATIONS

International standards are showing a willingness to provide a framework for the design and deployment of artificial intelligence (AI) throughout the entire lifecycle. A first approach has been structuring with the arrival of the specification on Frugal AI lead by AFNOR, the French organism for the standardization (see: "A benchmark for measuring and reducing the environmental impact of AI") and the French Government (see: Digital ecological footprint: standardization of frugal AI).

AI as part of a digital service or a product can already rely on existing robust standards (e.g., GHG Protocol, ISO/IEC on datacenters and software systems, the environmental assessment of products and services proposed by the ITU, etc.). To assess the environmental impact of digital services, the current standards use as references the ITU-T L.1480 "Enabling the Net Zero transition: Assessing how the use of information and communication technology solutions impact greenhouse gas emissions of other sectors", the ISO

14040.2006 “Environmental management - Life cycle assessment — Principles and framework” and the ITU-T L.1410 “Methodology for environmental life cycle assessments of information and communication technology goods, networks and services”.

However, approaches need to be harmonized to facilitate transparency and provide a common framework for assessing artificial intelligence.

- The first challenge is to define the scope of the calculations to be considered. There seems to be a consensus among experts on the life-cycle approach (from design to the end of life of artificial intelligence), but other movements want to go further (and for good reasons) by considering the indirect impacts and rebound effects generated by the products and the services that integrate AI.

- The second challenge will be to choose the right indicators to measure the environmental impact of artificial intelligence, to go beyond carbon and take into account consumption of water, equipment, etc.

Standardization remains a challenge, given the rapid pace at which AI technology is evolving, and the difficulty of mitigating the environmental impact of AI or AI systems involved in the development of technical solutions.

11. TOWARD FRUGAL AI INSPIRED BY NATURE

It is a striking fact that many of the basic behaviours requiring few efforts to animals are challenging to realize with current AI. These behaviours have been selected by millions of years of evolution to ensure animal survival, requiring them to solve as early as possible the so-called *“four Fs”*, namely feeding, fighting, fleeing, and mating. Although these behaviours may be learned and acquired by animals during their lifetime, it turns out that many of them are innate or are learned extremely quickly. This suggests that these innate mechanisms are wired up in the nervous system. However, simple calculations show that for animals with a large brain, DNA is not large enough to store all information about the nervous system connectivity [236]. Clearly, a larger brain allows the creation of new areas that don’t exist in a smaller brain, which can be recruited for the emergence of new behaviours or skills.

It seems, however, that for a given common cognitive task, the larger brains have a great deal of circuit redundancy, which ensures robustness and probably better discrimination between signals from sensory sensors. It is this redundancy, rather than the creation of new circuitry, that seems to be the main factor in the differences between larger and smaller brains [46]. Insects have much smaller brains than humans. They, however, often possess a very wide range of different behaviours, and are capable of complex learning (decisions, number evaluation, calculations, evaluation of time intervals time intervals, abstract comprehension, etc.), all at a very low energy cost [37]. For example, for a fruit fly (*drosophila melanogaster*) with an average weight of 1mg, the *total* metabolism requires around 0.1mW. In fact, it appears [46] that many of the cognitive tasks performed by insects require very few neurons and that brain size is not a reliable indicator of the diversity of cognitive behaviour. Beyond energy and structural aspects, numerous studies show that the creation of associative memory in insects’ brain is extremely fast and requires few training, exhibiting a form of a *few-shot learning* [181].

The combination low energy cost, circuitry of small size, and few-shot learning makes the brain of animals, and in particular of insects, particularly attractive as a source of inspiration for the design of frugal AI. Inspiration from general knowledge about brain structure has already a long history. Back to the seminal paper of W. S. McCulloch and W. Pitts in 1943 [154], the first neural networks were directly inspired by brain organization. Convolutional neural networks (CNN), now widely used in current AI models, are also inspired by the structure of the visual cortex of cats [82]. More recently, inspiration from the visual system of the dragonfly has been used toward the design of missile guidance and interception [42; 41]. Cerebellum inspired spiking neural networks are used in robotics for the control of articulation of unstable robots [175] or for multitask models for pattern classification and robotic trajectory prediction [215]. Moth and Drosophila’s olfactory circuits have been used to design image [65; 197] classification neural networks. Leveraging brain capabilities for frugal AI requires, however, deeper knowledge of its structural organization.

These models are based on the *functional connectome*, i.e., the connections between various *regions* of the brain. Leveraging brain capabilities for frugal AI requires, however more deeper knowledge on its structural organization given by the *neural connectome*, the wiring map at the neuron level. Until recently, connectomes of organisms were only partially known. The first complete connectomes were only characterized in the last decade for the roundworm *Caenorhabditis elegans* (302 neurons, 7000 synapses) initially available in 1989 [233] and revised in 2019 [53], for the tadpole larva of *Ciona intestinalis* (177 neurons, 6618 synases) [190] in 2016, for the segmented sea worm *Platynereis dumerilii* larva (1500 neurons, 25509 synapses) [213] in 2020, and for the drosophila larva (3016 neurons, 548000 synapses) [223] in 2023. Finally, in 2024 the full connectome of adult female Drosophila (139255 neurons, $5 \cdot 10^7$ synapses) has been reported [70]. In addition, several sub-circuits of these connectomes and their biological functions have already been identified. This is, for instance, the case for the regions associated with memory [136], its visual [204] and olfactory [193] systems, or its ellipsoidal body playing the role of a “compass” [106]. Overall, this detailed knowledge provides avenues for the design of frugal AI networks.

12. AI EMBEDDED ON DEVICES

This chapter presents basic information about dedicated hardware used in AI calculations: their types, characteristics, basic parameters, and usage scenarios.

12.1 Current State of Hardware for Frugal AI

The current state of frugal AI hardware focuses on solutions that combine computing power, cost-effectiveness, and

energy efficiency. Hardware has seen significant advances driven by the need to democratize AI beyond expensive, power-hungry systems like NVIDIA's H100 or Cerebras' WSE-2. The rise of edge computing has driven the development of low-cost neural processing units (NPUs), such as Qualcomm's Hexagon NPU in Snapdragon chipsets and AMD's Ryzen AI Engine in low-cost laptops, enabling AI model inference directly on the device with ultra-low power consumption. Companies like Google have shrunk the size of their Edge TPU to make it usable in more affordable devices like their Pixel phones, while startups like Groq and D-Matrix are introducing new designs, such as the Tensor Streaming Processor and in-memory computing chips, that maximize cost-to-performance ratios. Open-source hardware initiatives, like RISC-V-based AI accelerators, are also gaining traction, offering customizable, low-cost alternatives to proprietary ASIC solutions. Meanwhile, energy-efficient photonic chips from Lightmatter and neuromorphic processors like Intel's Loihi 2, whom remain in early adoption stages, but promise to further reduce operational costs. Overall, these developments signal a shift toward frugal AI hardware that balances performance and affordability, making AI more accessible on many more devices at much lower cost.

12.2 Dedicated AI Hardware

12.2.1 Overview of dedicated AI hardware

Traditional general-purpose processors (CPUs) are often incapable of handling the massive computational loads required by modern AI applications. This has led to the adaptation of already existing or the development of new types of devices supporting AI tasks, which may be called AI accelerators.

AI accelerators are specialized hardware designed to speed up the computation processes needed for artificial intelligence (AI) and machine learning (ML) tasks. These devices are optimized to handle the massive parallelism and high-performance demands of AI workloads, such as training deep neural networks, running inference tasks, and processing large datasets.

Computational models: There are two primary models for AI computing: cloud-based and edge, each offering distinct advantages and trade-offs. Understanding these models is essential in choosing the right solution for specific use cases, particularly in the context of frugal AI, where efficiency, cost, and performance are crucial.

AI accelerators for cloud computing and edge computing are often designed with different priorities and use cases in mind, so they typically look different in terms of form factor, performance characteristics, and power consumption (see Table 4).

Types of AI accelerators: We can distinguish several types of these devices:

- **Graphics Processing Units (GPUs):** originally designed for graphics rendering, GPUs are highly parallel processors that are well-suited for deep learning tasks, particularly for training neural networks.
- **Tensor Processing Units (TPUs):** developed by Google, TPUs are application-specific integrated circuits (ASICs) designed to accelerate tensor processing. TPUs offer high efficiency and are tailored for workloads using Google's TensorFlow framework.
- **Field-Programmable Gate Arrays (FPGAs):** FPGAs are configurable hardware that can be customized to optimize specific AI algorithms. They offer flexibility for fine-tuning[20] AI applications but may not reach the same level of performance as GPUs or TPUs in certain tasks.
- **Application-Specific Integrated Circuits (ASICs):** these are custom-designed chips built specifically for AI workloads. They provide excellent performance but are limited to specific tasks.
- **Neural Processing Units (NPUs):** NPUs are specialized hardware designed specifically for accelerating neural network-based algorithms. They are found in some modern smartphones and embedded systems.
- **Language Processing Unit (LPU):** LPU is a proprietary and specialized chip developed by the Groq company. It is designed to handle the unique speed and memory demands of LLMs – tasks that are sequential by nature rather than parallel.
- **Digital Signal Processors (DSPs):** while not as specialized as others, DSPs can accelerate certain signal processing tasks related to AI, such as audio and image processing, with lower power consumption.

AI accelerators play a critical role in the evolution of AI technologies, making complex computations more efficient, faster, and cost-effective, which is essential for the rapid progress of AI applications across various industries.
The following table compares basic features of different types of AI accelerators (their architectures).

12.2.2 AI accelerators in embedded systems (for Frugal AI)

This chapter focuses on AI accelerators used in embedded systems in the context of "Frugal AI". We discuss requirements imposed on this type of equipment, types of devices, their characteristics, as well as their advantages, disadvantages, and challenges.
While AI accelerators such as GPUs, TPUs, NPUs, and FPGAs have traditionally been used in high-performance data centers or cloud-based systems, the shift towards edge AI and frugal AI solutions is reshaping the landscape. Frugal AI refers to the application of AI technologies in environments with constraints such as limited power resources, low-cost hardware, small form factors, and low-latency requirements. This shift demands the use of low-power, cost-effective, and efficient AI accelerators capable of performing high-speed computations without compromising energy consumption or operational costs.

AI accelerators can be very useful in the context of Frugal AI, especially in environments with limited computing power or budget. The concept of Frugal AI often focuses

[20] See Section 14.2 for a definition of fine-tuning

Feature	Cloud computing	Edge computing
Form factor and hardware design	- usually high-performance, large-scale devices like GPUs, TPUs, or ASICs (housed in data centers), - designed to handle the heavy lifting of AI tasks such as training deep neural networks or processing large datasets in real-time across many users. - can be rack-mounted or part of large-scale server systems, and are typically more power-hungry, as they can rely on high power and cooling systems provided by the data center.	- typically compact, energy-efficient, and designed for low-power environments. They need to be small enough to fit in devices like smartphones, IoT devices, drones, autonomous vehicles, and embedded systems. - often designed to provide AI capabilities directly on the device without relying on cloud computing, enabling real-time processing and low latency in scenarios like real-time video processing, voice assistants, or autonomous decision-making.
Performance characteristics	- Optimized for maximum computational power, which is necessary for training large models and performing complex computations that require extensive parallel processing. - Typically handle tasks like large-scale machine learning training, processing large datasets, and executing high-throughput operations. The performance (measured in terms of teraflops, for example) is much higher compared to edge accelerators. - Have virtually no constraints on power or thermal limits, as they are typically in large data centers with access to robust cooling systems.	- Optimized for lower power consumption while still delivering sufficient performance to handle real-time AI inference tasks. They are designed to run pre-trained models (inference), rather than training new models. - Performance is usually lower compared to cloud accelerators, but the focus is on balancing speed, power efficiency, and small size. - The goal is to perform local processing to reduce the need for constant communication with the cloud, improving latency and privacy.
Power consumption	- Generally not constrained by power limitations, as they reside in data centers with access to ample power and dedicated cooling solutions. They can consume a significant amount of energy due to their high-performance design.	- Power efficiency is a critical factor here. These accelerators are designed to operate on devices with limited power supply, like smartphones, wearables, or battery-powered IoT devices. Power consumption must be minimized without sacrificing too much performance.
Use cases	- Training large-scale AI models (e.g., training deep neural networks for natural language processing, image recognition, etc.). - High-volume AI inference for tasks like recommendation systems, fraud detection, and serving multiple clients with complex models. - Examples: data centers processing AI for online services, such as search engines, recommendation engines, and advanced analytics.	- Real-time inference on localized devices, enabling low-latency processing without waiting for cloud communication. - Common edge computing tasks include autonomous vehicles, smart cameras, IoT sensors, voice assistants, and smartphones. - Examples: on-device image recognition for surveillance cameras, facial recognition on smartphones, voice-to-text on smart speakers, and real-time decision-making in drones or robots.
Connectivity and latency	- Rely on high-speed internet and cloud infrastructure for communication. This introduces latency due to the need for data transfer between the edge device and the cloud, especially in remote or poorly connected areas.	- Aim to minimize or eliminate latency by processing data directly on the device, which can be crucial for time-sensitive tasks (e.g., autonomous driving, real-time medical diagnostics). - Data is processed locally without the need for an internet connection, ensuring that decisions can be made instantaneously.
Cost	- The cost of using cloud-based AI accelerators is typically usage-based and can be expensive for extensive tasks like model training or large-scale data processing, though it offers scalability and flexibility. - Costs can include cloud service subscriptions, data transfer, and storage fees.	- Typically more affordable in terms of upfront costs, as they are embedded in consumer devices or dedicated hardware for specific applications. - While the initial cost may be lower, managing a large-scale network of edge devices could still involve infrastructure management and maintenance costs.

Table 4: Computational models

on building AI models and solutions that achieve significant results with minimal resources, which is especially important in settings like emerging markets, low-cost devices, or resource-constrained environments.

Table 5 describes how AI accelerators align with and enhance Frugal AI.

12.2.3 Types of AI accelerators in/for embedded systems

AI accelerators for embedded systems come in various forms, including low-power GPUs, NPUs, FPGAs, and ASICs, each offering unique advantages depending on the specific application requirements. What sets these accelerators apart is their ability to deliver high compute performance while maintaining low power consumption and occupying minimal space: two critical factors in embedded applications.

Low-power GPUs: Low-power GPUs are designed specifically for embedded systems, mobile and IoT devices, smart cameras, drones and edge computing where energy efficiency is crucial. They deliver a balance between performance and power efficiency, making them suitable for battery-operated devices and energy-constrained applications.

Examples of this type of device are:

- NVIDIA Jetson Series (Jetson Nano, Jetson Xavier NX) [165]
- ARM Mail GPUs (Mali-G52, Mali-G76, Mali-G57) [15]
- Qualcomm Adreno GPUs (Adreno 620, Adreno 660) [179]
- Intel Integrated Graphics (Iris Plus, UHD Graphics)
- AMD Radeon RX 500 Series (low-power models)
- Imagination Technologies PowerVR Series (GM9446, Series8XE) [56]

Feature	NPU	GPU	TPU	FPGA	CPU	ASIC
Optimization Target	Deep learning inference (CNNs, RNNs, Transformers)	Parallel processing (Graphics, AI, HPC)	Tensor operations (ML training & inference)	Custom AI workloads	General-purpose processing	Fixed AI models (optimized for efficiency)
Processing Units	Specialized MAC (Multiply-Accumulate) Arrays, SIMD	Thousands of CUDA cores for parallelism	Large-scale matrix multipliers & systolic arrays	Reconfigurable logic gates	Few general-purpose cores	Custom AI logic circuits (non-reprogrammable)
Precision	Optimized for low-precision (INT8, FP16, BF16)	Supports FP32, FP16, INT8	Uses BF16, INT8 for efficiency	Programmable for various precisions	Typically FP32, FP64	Optimized for fixed precision (INT8, FP16)
Memory Access	Tightly coupled SRAM/DRAM for fast AI data access	High-bandwidth GDDR/VRAM for large models	High-bandwidth memory (HBM) for large tensor ops	Custom memory configurations	Uses caches & RAM for general computing	Custom memory architecture (on-chip & external memory support)
Power Efficiency	Very high (1-10 TOPS/W)	Moderate (0.1-1 TOPS/W)	High (5-10 TOPS/W)	Variable	Low for AI (not optimized)	Very high (>10 TOPS/W, but fixed function)
Flexibility	Fixed-function for AI	Programmable for AI, graphics, and compute	Fixed-function for deep learning	Highly flexible & reprogrammable	General-purpose, least optimized for AI	Fixed-function, cannot be reprogrammed
Latency	Ultra-low (real-time inference)	Moderate latency	Low latency (batch processing)	Varies (can be optimized)	High latency for AI	Ultra-low (dedicated for specific AI models)
Programming Complexity	Easy (pre-optimized AI frameworks)	Moderate (CUDA, OpenCL)	Moderate (TensorFlow XLA)	High (HDL, Verilog, VHDL)	Simple for general tasks, slow for AI	Low (hardwired AI logic, minimal software adaptation)
Use Cases	Edge AI, smartphones, IoT, AI cameras	AI training, HPC, gaming, ML inference	AI training & inference, cloud AI	Edge AI, IoT, custom applications	General computing, OS tasks	Dedicated AI tasks (speech, vision, data center AI, crypto mining, automotive AI)

Figure 8: AI Accelerators feature comparison

- VPU (Vision Processing Unit) by Intel Movidius. [110]

These low-power GPUs are suitable for applications in Frugal AI, as they make AI more accessible by reducing the cost and energy consumption needed to run AI models, especially in environments with limited resources.

Coral Edge TPU: Google Edge TPU is a specialized low-power AI accelerator designed for edge computing. It provides fast, efficient machine learning inference while consuming minimal power, making it ideal for IoT, embedded AI, and smart devices. Its key features are:

- **ultra-low power consumption:** ideal for battery-powered AI devices,
- **optimized for TensorFlow Lite:** fast and efficient inference for pre-trained models,
- **cost effectiveness:** a relatively low-cost solution for running AI models on edge devices,
- **affordable and scalable:** integrated into Coral Dev Boards, USB accelerators, and M.2 modules,
- **real-time AI at the edge:** no need for cloud processing, reducing latency and data transfer costs,
- **user-friendly:** easy to integrate with popular Raspberry Pi boards and other small devices.

Field-Programmable Gate Array (FPGA) AI accelerators: FPGAs are hardware devices that consist of an array of programmable logic blocks, which can be configured to execute custom operations. These devices are highly flexible and can be adapted to meet specific computational needs. The advantages of using FPGAs for AI acceleration are:

- **customizable processing pipelines:** they can be programmed to implement custom hardware accelerators for specific parts of an AI model,
- **energy efficiency:** they offer lower power consumption compared to GPUs and CPUs for specific workloads, i.e., a well-optimized FPGA can provide performance similar to GPUs but with much less power usage,
- **high throughput and parallelism:** the ability to perform multiple operations in parallel allows FPGAs to provide high throughput for AI workloads,
- **low latency:** they have a unique advantage when it comes to low-latency AI inference,
- **reconfigurability:** unlike specialized AI hardware accelerators like ASICs, FPGAs can be reconfigured to support new algorithms or updated models.

Features	Description of AI accelerators
Improved Efficiency with Limited Resources	They can perform AI tasks much faster than general-purpose CPUs, helping achieve better performance without needing large-scale, expensive infrastructure.
Cost-Effective AI Solutions	Allow for cost-effective solutions by providing specialized hardware that delivers high performance without requiring a significant investment. Becoming more common, enabling the deployment of AI in resource-constrained environments while keeping costs low.
Energy Efficiency for Sustainable AI	Designed to be more energy-efficient than general-purpose processors, which is critical when deploying on battery-operated devices or in areas with limited power resources. Remain sustainable and can be deployed at scale, even in environments where electricity costs are high or where access to power is limited (e.g., rural areas, developing countries).
Enabling Localized AI for Accessibility	Frugal AI often focuses on local processing (i.e., on-device AI), which ensures that AI applications are available even in remote areas with limited connectivity.
Scalability with Low-Cost AI Infrastructure	In many parts of the world, AI applications need to be deployed on a large scale but with limited resources. AI accelerators in smartphones, IoT devices, or embedded systems offer a way to scale AI solutions across many devices with minimal cost.

Table 5: AI accelerator features that boost Frugal AI.

There are also some challenges while using FPGAs for AI:

- **programming complexity:** one of the biggest challenges of using FPGAs is the programming complexity, because it requires knowledge of hardware description languages (HDL),
- **performance variability:** the performance depends heavily on a configuration of a particular task. Poor optimization can lead to suboptimal performance. As a result, performance tuning is essential, which can be time-consuming,
- **cost and availability:** they can be more expensive than GPUs for some use cases, particularly for mass deployment in cloud-based or consumer devices.

Here are several examples of FPGA AI accelerators: Xilinx Versal AI Core [9], Xilinx Vitis AI [10], Intel Altera [109], Achronix [54], AWS EC2 F1 instances [72].

ASICs for AI acceleration: ASICs are custom-designed hardware solutions optimized to perform specific tasks much faster and more efficiently than general-purpose processors (CPUs and GPUs). The key points of ASICs as AI accelerators are:

- **specialization:** ASICs are built for one particular job. By tailoring the hardware to a specific AI model or operation, ASICs are highly efficient at executing those tasks,
- **high performance:** they can achieve unmatched, processing many operations in parallel with minimal overhead,
- **low power consumption:** can be extremely power-efficient because the hardware is tailored to the task at hand,
- **fixed functionality:** that means they are incredibly efficient at doing what they are designed to do,
- **cost-effectiveness at scale:** while ASICs can be expensive to develop initially, they become extremely cost-effective at scale,
- **compact form factor:** ASICs can be designed to have a very small form factor, which allows them to be integrated into compact devices.

Despite these advantages, ASICs also meet some challenges:

- **lack of flexibility:** ASICs are fixed-function devices, meaning that once designed, they cannot be reprogrammed or repurposed for other tasks,
- **high development cost:** designing and manufacturing an ASIC is a costly and time-consuming process, typically requiring millions of dollars in research and development, especially for custom-designed hardware,
- **initial investment:** the upfront cost to develop and produce an ASIC is significant,
- **limited customization after production:** once an ASIC is produced, any changes to the hardware require the creation of a new version.

Examples of ASIC AI accelerators are: Google TPU [50], Apple's Neural Engine (ANE), Huawei Ascend [55], Intel Nervana NNP (discontinued in favor of development of Habana Labs' chips) [111].

Neural Processing Unit(s): Neural processing units (NPUs) are specialized computer microprocessors designed to mimic the processing function of the human brain. They are typically used within heterogeneous computing architectures that combine multiple processors, e.g., CPUs and GPUs on a single semiconductor microchip known as a system-on-chip (SoC).

By integrating a dedicated NPU, manufacturers are able to offer on-device generative AI apps capable of processing AI applications, AI workloads, and machine learning algorithms in real-time with relatively low power consumption and high throughput.

The following list presents NPUs' key features:

- **parallel processing:** NPUs can break down larger problems into components for multitasking problem solving,
- **low precision arithmetic:** NPUs often support 8-bit (or lower) operations to reduce computational complexity and increase energy efficiency,
- **high-bandwidth memory:** high-bandwidth memory on-chip feature to efficiently perform AI processing tasks requiring large datasets,
- **hardware acceleration:** incorporation of hardware acceleration techniques such as systolic array architectures or improved tensor processing.

Examples of NPU AI accelerators are: Rockchip RK3399Pro [186], MediaTek Dimensity NPU [156], Khadas Vim3 [121], Huawei Ascend CPUs [55], Arm Cortex-M55 [14], Arm Ethos-N78 [13].

12.3 Future Trends in Hardware for Frugal AI

Next-Generation Chips:

- Predictions on how processors will evolve to better support AI tasks with minimal resources.
- Focus on energy efficiency, speed, and computational power.

Emerging Technologies: Emerging technologies can help stem the growing resource needs of today's AIs by bringing new ways of thinking about and implementing computing algorithms. Among these emerging technologies, quantum and neuromorphic computing offer a seemingly more sustainable alternative to "classical" deep learning.

- *Quantum computing*: leveraging quantum superposition and entanglement phenomena offers an approach to computing where all possible results of a given calculation can be done in a single step, whereas they should be treated sequentially with classical computers. This should allow tremendous speed-up of computation, allowing to tackle problems that are practically impossible to address by using classical computing. Numerous research works aim at rethinking machine learning in the light of quantum computing [237]. Another appealing property of quantum computing is related to the fact that quantum computing systems use energy in a very different way than classical computers. Quantum computing is very low in terms of energy consumption. The main energy cost in quantum computer systems is due the cryogenic cooling [216], since it must operate at low temperature (close to the near absolute zero). If for classical computers, the energy cost scales roughly linearly with computational power, increasing the number of qubits by several orders does not necessarily require increasing the cooling energy. As a consequence, the energy cost of a quantum system scales much more slowly with respect to computation capabilities than classical systems.
- *Neuromorphic computing* can be seen as the association of *spiking neural networks* (SNN) [164; 130] and efficient devices like *memristors* [226], both drawing inspiration from brains. In contrast to "classical" neural networks (DNN - Deep Neural Networks), SNNs are event-driven neurons, emitting a spike (an impulsion) when their internal potential, driven by incoming spikes, reaches a certain value. A spiking neuron needs energy only during a spike emission. Altogether, a spiking neuron constitutes both a memory and a computation unit. This allows breaking the Von Neumann bottleneck by drastically reducing the energy required to transfer data and speeding up data processing. At a low level, memristors are used to implement spiking neurons in an extremely energy-efficient way. Due to their dynamical behaviour, SNNs are also particularly adapted to real-time analysis (e.g., [214]). Methods allowing transformations from DNN to SNN are available in [36] and its references. Many architectures inspired by the DNN have been designed using SNN-like convolutional layers [228] or even attention layers and transformers [138].

However, the recent progress in neurology and in the identification of neural circuits in brains (see Section 11) may open many new opportunities to draw inspiration from the small and efficient substructures found in real neural systems.

Custom AI Chips:

- Trend towards ASICs designed specifically for AI in embedded systems.
- Companies like Tenstorrent, Mythic, and Hailo with their unique offerings.

13. AI OPTIMIZATIONS

The great success of Deep Learning methods [131] in numerous domains comes with the two major drawbacks: availability of computing power and of the vast quantity of training data. Frugal approaches are diametrically opposed to Deep Learning methods. In this section, we review some optimization approaches that have been proposed in the literature to enforce frugality in Deep Learning. Model compression techniques (see 13.1) are used to decrease the memory footprint and computational complexity of deep learning models. Hardware optimization techniques (see 13.2) aim at defining dedicated hardware solutions in order to enhance computational efficiency, reduce latency, and minimize energy consumption, whereas deployment techniques (see 13.4) address the optimization of resource deployment. Algorithmic optimization techniques (see 13.3) tackle the learning process and are used when training and inference tasks have limited compute resources. Finally, data-efficiency methods (see 13.5) are crucial, especially if datasets are non-accessible (rare, expensive, or private).

13.1 Model Compression Techniques

Considering the cost of AI systems (see section 7) with deep neural-based models, optimizing the model itself may help decrease the infrastructure cost, the training or retraining and inference costs, or even the deployment cost. Model compression techniques are an umbrella under which several different approaches are undertaken in order to reduce these costs. These techniques aim at decreasing one or several of the technical metrics given in Figure 9 while simultaneously maintaining the model performances (accuracy, precision, etc). These metrics are, however, not independent. For instance, decreasing the FLOPS (floating point operations, roughly the number of additions and multiplications), evaluating the computational complexity of the model may increase the number of costly memory accesses, increasing the backward and forward latency.

Over the last decades, many model compression strategies have been proposed in the literature and good general surveys are available like for instance [230; 158; 150] or [157]. Surveys are also available dedicated compression methods

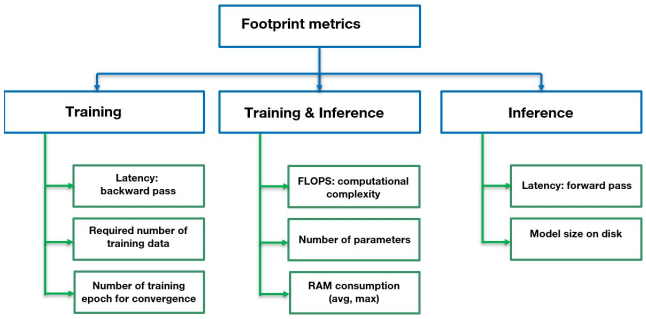

Figure 9: Main model metrics addressed by model optimization techniques for deep neural models.

applied to AI models with specific structures like Convolutional Neural Networks [133], Transformers [206] or with specific tasks like image classification [188] or large language machine [229; 244]. The main available strategies can be categorized as:

- **Quantization**: In a typical deep neural network model, weights, gradients, and activations are typically represented as 32-bit floating point numbers, a precision level resulting in high power consumption and high memory resource requirement. Quantization methods aim at replacing these high-precision values by more compact ones (16-bits, 8-bits, ternary or binary), reducing memory footprint and/or by more efficient ones, e.g. logarithmic quantization allowing replacing costly multiplications by bitshift operations [143]. Surveys of these techniques can be found in [58; 85] or [230].
- **Pruning**: Removing unimportant neurons and connections (*unstructured pruning*) or even full substructures (e.f. channels or filters in CNN, attention heads in transformers) or layers (*structured pruning*) in order to decrease the memory footprint and the computational complexity of a model. Accounts on pruning methods can be found in [45; 210] or [97] for CNN-based models.
- **Low-Rank Approximation**: Approximating high-rank matrices with low-rank counterparts to reduce memory footprint and/or computational complexity. These methods typically leverage singular value decomposition, matrix factorization, or tensor decomposition. Surveys of these approaches can be found in [230; 171] or [172].
- **Knowledge Distillation**: Using a large and complex model (the *teacher*) to train a smaller and simpler one (the *student*). The distillation process can be performed during the training of the teacher (*online* distillation) or using the pre-trained teacher (*offline* distillation). Good accounts of this type of method can be found in [162] or [230].
- **Neural Architecture Search (NAS)**: For a given task and a given dataset, use an algorithm to automate the search of optimally compact and efficient artificial neural networks performing as well or even outperforming hand-crafted neural network architectures. Recent surveys can be found in [47; 74; 220].

Although these methods are the most commonly used, other approaches are also proposed. For instance, in order to minimize the memory footprint of large weight matrices, *sparse representation* like **weight sharing** aims at transforming many similar parameters with a single connection into a single weight with multiple connections [161]. Other approaches referred to as **lightweight design** propose to replace standard structures with simpler and more efficient ones. For instance, dilated convolution [235]. Furthermore, all these previous methods can be used alone, in combinations, or associated with other ones. For instance, regularization techniques [205] can be used to enforce sparsity in model parameters in combination with pruning.

13.2 Hardware Optimization Techniques

Hardware optimization techniques in artificial intelligence (AI) are pivotal in enhancing computational efficiency, reducing latency, and minimizing energy consumption. These techniques encompass various strategies, each contributing uniquely to the performance of AI systems.

13.2.1 Specialized Hardware Accelerators:

The development of hardware accelerators, such as **Graphics Processing Units (GPUs)**, **Tensor Processing Units (TPUs)**, and **Field-Programmable Gate Arrays (FPGAs)**, has been instrumental in optimizing AI workloads. These accelerators are designed to handle the parallel processing demands of AI algorithms, thereby improving throughput and energy efficiency. For instance, FPGAs offer customizable hardware solutions that can be tailored for specific AI applications, providing a balance between performance and flexibility. [196], [26]

In certain high-performance or high-efficiency use cases, the co-design of hardware and software can encompass the creation of dedicated hardware accelerators (Application Specific Integrated Circuits – ASICs) for the particular AI model. By tailoring software algorithms to leverage specific hardware features, and vice versa, this technique achieves efficient execution of AI tasks. For example, optimizing models for specific hardware platforms, such as Intel Xeon processors, can lead to significant performance gains [17]. This approach is the most efficient but entails a high degree of investment and technical knowledge.

13.2.1.1 Application-Specific Integrated Circuits (ASICs).

[141] are custom-designed integrated circuits tailored for specific applications, offering optimized performance, reduced power consumption, and enhanced efficiency compared to general-purpose hardware. They are usually created from the ground up, based on the specific needs of the application they are intended for. On 10, existing types of ASICs [38] are illustrated. Examples of ASICs span various domains, including:

- **Telecommunications:** ASICs are employed in network routers and switches to handle specific protocols

and data processing tasks, enabling high-speed data transmission and efficient network traffic management.

- **Consumer Electronics:** Devices such as smartphones, digital cameras, and gaming consoles utilize ASICs to manage specific functions like signal processing, power management, and audio encoding/decoding, contributing to enhanced performance and reduced power consumption.
- **Automotive Industry:** Modern vehicles incorporate ASICs for various applications, including engine control units, airbag deployment systems, and advanced driver-assistance systems (ADAS), ensuring real-time processing and increased reliability

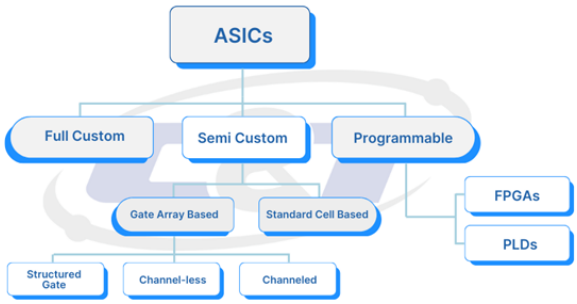

Figure 10: Types of ASICs (figure from [38])

13.2.2 *Advanced Matrix Extensions (AMX):*

Introduced by Intel, AMX is an extension to the x86 instruction set architecture designed to accelerate matrix operations, which are fundamental in AI and machine learning workloads. AMX enhances computational efficiency by introducing two-dimensional registers and specialized accelerators for matrix multiplication, thereby improving performance in AI applications. [1]

13.2.3 *Hardware-Based Memory Optimization Techniques:*

- **High-Bandwidth Memory (HBM):** Specialized memory like HBM2 and HBM3 (used in NVIDIA A100, AMD MI300) integrates memory closely with processing units, providing increased bandwidth and reduces memory bottlenecks. This proximity allows for faster data transfer rates, essential for AI tasks that require rapid access to large datasets. Implementations of HBM in AI accelerators have demonstrated significant performance improvements in deep learning applications. [123]
- **On-Chip Memory Optimization:** AI accelerators like TPUs, FPGAs, and ASICs reduce reliance on external memory by using on-chip SRAM or eDRAM, decreasing memory access latency. [22]
- **Memory Hierarchy Optimization:** Advanced caching mechanisms and memory prefetching techniques (e.g., L1/L2 cache optimizations in AI chips) improve data access speeds. [211]
- **Specialized Memory Architectures:** Custom memory designs, such as those utilizing metal-oxide combinations in RRAM, offer non-volatile storage solutions with high endurance and speed. These characteristics are beneficial for AI applications requiring persistent storage and rapid data retrieval. Research into metal-oxide RRAM has highlighted its potential in neuromorphic computing and AI hardware acceleration.[95]
- **Processing-in-Memory (PIM):** Emerging PIM architectures integrate processing units directly within memory modules, minimizing data movement overhead. Recent advances in PIM have shown promise in optimizing analogue AI computations, particularly through the use of resistive random-access memory (RRAM) technologies. [139]
- **Hardware-Assisted Mixed Precision Support:** Modern GPUs (e.g., NVIDIA Tensor Cores) and AI ASICs (e.g., Google's TPUs) provide native support for lower-precision computations (FP16, INT8) to optimize memory usage. [184]
- **Accelerator-Driven Data Arrangement:** Optimizing data placement and access patterns in memory can significantly reduce runtime for AI models. Techniques that align data organization with the architecture of hardware accelerators have been shown to minimize off-chip data access, thereby enhancing performance in transformer-based models. [11]

For a comprehensive understanding of these hardware optimization techniques and their applications, several literature reviews provide in-depth analyses. [5], [142] , [137]. These resources collectively elucidate the critical role of hardware optimization in advancing AI capabilities, particularly in environments with stringent resource constraints.

13.3 Algorithmic Optimization Techniques

There are two main algorithm optimization approaches: increasing the efficiency of training or inference. The major training optimization methods are:

- **Distributed Learning** over decentralized hardware has become an important challenge with the emergence of powerful personalized equipment, capable to train and/or execute various applications *on-the-chip* (Internet-of-Things or smartphone devices). We distinguish two major approaches: Federated and Split Learning. *Federated Learning* [128], [127], [34] has emerged as a key solution to reduce the need for centralized data gathering and training. This collaborative and iterative approach builds a common global model. The model benefits from local knowledge learned on private data, without sharing data with third parties. *Split learning* methods [92] are deployed when data labels are delocalized from data gathering equipment, or if the capacity of the training device is not sufficient to execute a single iteration. Recently, the hybrid methods [140] have emerged. They benefit jointly from the advantages of split and federated learning.
- **Meta-learning** methods [208], [102] belong to the class of learning algorithms whose performance increases not only with the number of training samples, but

also with the number of (potentially related) learning tasks. This concept (*learning to learn*) is similar to the animal learning process (learning biases and generalizations, given a few examples), which improves the data and computation efficiency.

- **Reinforcement Learning** (LR) [201], [117] maximizes the total reward of the *agent* over interactions with uncertain and complex environment. The two threads represent the *trial-and-error* learning system and optimal control. In certain cases, it is possible to simplify the calculus load or minimize the latency or energy consumption by splitting a single agent into multiple agents [145], or by their spatial distribution. The very recent developments in GenAI in combination with the emergence of Agentic AI [71] lean on the RL approaches to minimize the overall calculus energy. The open question remains if the former two methods cost less than simpler but equivalent Machine Learning approaches that can not generalize to multiple tasks.

- **Self-supervised learning** (SSL) grasps the dependencies between its inputs from a large volume of unlabelled instances. This is one of the human-level intelligence factors, and its principles are used to train early NN networks [23], [100]. SSL learns discriminative features by automatically generating pseudo-labels. One way to create these labels is by data-augmentation: building the transformations of a single sample (so-called *dictionary*) and aligning it to similar or dissimilar samples.

  There are four classes of the SSL [20]: Deep Metric Learning (DML), Self-Distillation, Canonical Correlation Analysis (CCA) and Masked Image Modeling (MIM). The DML methods train networks to distinguish sample pairs that are alike in the embedding, and some also perform mining of the similar pairs present in the original dataset. The class of Self-Distillation algorithms learns the predictor to correctly map the outputs of the two encoders, which were fed by the two similar (transformations of the single input) or dissimilar samples. One way to prevent the *predictor collapse* (prediction of the constant) is to use two predictor networks, *student* and *teacher*. They are updated throughout the training by using gradient descent (student) and moving average-based weight updates of the student network (teacher). The CCA is a family of methods that analyses the cross-covariance matrix of variables to infer their relations. For multivariate and nonlinear CCA, one popular way to do this is to jointly learn parameters of the two networks with maximally correlated outputs.

- **Transfer Learning** (TR) [173] promotes the lifelong machine learning knowledge re-usage to minimize the latency and energy used for training. In general, the transfer of knowledge towards the current task considers already gathered datasets or models trained prior to the current task. Data-based approaches are focusing on transformations between datasets (feature-relations, distributions, etc.). Model-based TR initializes the training model with the existing one (or its adapted version), which is often trained in domains, tasks, and distributions that are different from the current task. Based on the similarity of the feature space [219], TR can be split into homogeneous (domain differences are modelled by bias or conditional distribution corrections) and heterogeneous TR. There are globally four TR [173]: instance- (heuristic or hypothesis based instance-weighting methods), feature- (transformation of the original feature set towards symmetric or asymmetric feature representations: augmentation, reduction or alignment of distribution differences), parameter- (model and/or parameter transfer of knowledge) or relational-based methods (transfer of the *source-target* relationship rules: spatial or geometric structure, statistics, etc.).

- **Multi-task** [40] is an inductive transfer learning approach that trains a common model over different tasks. The intuition behind this is that the generalization of the model improves even if training tasks are not related. Its training cost is smaller than that of a cumulated sum of *per-task* training. The learning complexity of multi-task algorithms varies, ranging from k-nearest neighbours (sharing the clustering structure [114]), decision trees [108] (feature subset share), towards backpropagation neural networks (multiple outputs that share one fully connected hidden layer, for example). Today, distributed and asynchronous variants of multi-task learning boost its usage. Moreover, trained models deployable to continual or active learning may outperform approaches that do not use transfer learning [183].

- **Instance-based** methods [2] do not train any model, but rather use the available dataset for prediction on new data. It is efficient, but in general less accurate compared to algorithms based on model training. It is used in cases It is often used in pattern recognition or anomaly detection fields.

The above list of training techniques that may improve efficiency is not exhaustive. The final choice of the algorithm depends on a set of specific parameters of a use case (energy consumption, hardware, topology, etc.). Other efficient techniques exist, such as weakly-supervised or incremental learning.
The outcome of the training is a model that is further deployed on one or more types of equipment for *inference* (i.e., detection, classification, prediction, etc.). The major inference optimization methods are:

- **Distributed inference** allows for deployment of the trained models on edge-like equipment to achieve quicker response times, reduced bandwidth costs, and enhanced data privacy.

- **Model compression and approximation**: it is possible to use approximate solutions (i.e., quantized, pruned models) to reduce the overall computational complexity.

- **Other classes of inference accelerations**: early exit of inference, inference cache, or model-specific inference accelerations (CNN, RNN, Transformer) [7].

13.4 Deployment Optimization Techniques

13.4.1 Efficient serving strategies

- **Serverless Computing:** Serverless architectures enable dynamic resource allocation, allowing AI models to scale efficiently based on demand. This approach reduces operational costs and simplifies deployment, particularly in high-volume applications. [96]
- **Cloud-Based Deployment:** Utilizing cloud platforms for AI deployment offers scalability, flexibility, and access to powerful tools and infrastructure, which are built to be energy efficient. Best practices include selecting the appropriate cloud platform, optimizing data storage and management, implementing robust security measures, and monitoring performance to ensure cost-effectiveness and efficiency [174]
- **Multi-tier serving:** Deploying lightweight models on edge devices for rapid responses, while utilizing more comprehensive models on the cloud for high precision when necessary, is suitable for applications that balance speed and accuracy, such as speech assistants and mobile AI. [3]

13.4.2 Parallelization, Distributed Training & Inference

- **Model Parallelism:** Dividing a model across multiple GPUs or TPUs is beneficial for very large models. [242]
- **Data Parallelism:** Distributing input data across multiple processing units facilitates faster inference. [195]
- **Edge-Cloud Hybrid Inference** (similar to Multi-tier serving & Load Balancing Across Distributed Systems): Offloading intensive computations to the cloud while maintaining lightweight operations at the edge optimizes performance and resource utilization. [240]

13.4.3 Scaling strategies

- **Adaptive Computation Scheduling:** Dynamically allocating computational resources based on runtime conditions, such as prioritizing critical tasks or adjusting inference frequency, thereby optimizing latency and energy use. [28]
- **Load Balancing Across Distributed Systems** (similar to Multi-tier serving & Edge-Cloud Hybrid Inference): Ensuring efficient resource utilization in multi-device or cloud-edge deployments by distributing inference tasks according to device capacity and network conditions. [118]
- **Context-Aware Inference:** Leveraging environmental or user-specific cues to selectively activate model components, reducing unnecessary computation. [207]

13.4.4 Graph substitutions

Each substitution replaces a sub-graph matching a specific pattern with a new sub-graph that computes the same result. What is worth emphasizing is that the architecture of the model does not change as a result of these operations. For example, operator fusion combines multiple operators (e.g., BatchNorm, ReLU, and Conv) into a single kernel, reducing memory access overhead and enhancing performance during inference. [199]. [76], [116]

13.4.5 Examples of deployment optimization tools and frameworks

They usually mix different techniques, described in the subsections above. These are, for example:

- **TVM (Apache TVM):** An end-to-end deep learning compiler that optimizes model execution for different hardware targets (CPU, GPU, FPGA, and microcontrollers). [12]
- **XLA (Accelerated Linear Algebra):** A domain-specific compiler for optimizing TensorFlow and JAX models. [169]
- **OpenVINO:** provides graph optimizations, operator fusion, and low-level execution improvements similar to other compiler-based tools. It targets specific Intel accelerators (e.g., CPUs, GPUs, FPGAs, VPUs). [168]
- **TensorRT (Nvidia):** Converts and optimizes deep learning models for high-performance inference on NVidia GPUs. [166]
- **ONNX Runtime:** is a cross-platform machine-learning model accelerator [167]

13.5 Data efficiency methods

The choice of the frugal algorithm should take into account the specificities of input data (i.e., availability of labels for learning, volumes: large/rare dataset, structure, etc.), its properties (modality, correlations, etc.) and the final usage (single, multi-task, future transfer learning, etc.).

- **Online Learning**: This class of algorithms [101] learns incrementally from new data. This allows adaptations in evolving environments without revisiting past data (for example, change of data distributions).
- **Data augmentation**: Data storage capacity is sometimes poor. Data augmentation methods increase the number of samples used in training, given a modest dataset size. Particular methods range from generative augmentation, feature-space augmentation, unsupervised augmentation, or basic transformation functions, see [217] and references within. Several categorizations are possible, for example, based on the number of samples used for a new sample generation (individual, multiple, or population data augmentation) or based on data-modality (value-, structure- or value-structure data augmentation).
- **Knowledge sharing** (i.e., meta learning [103], lifelong learning [189], multi-task learning)

- **Non-supervised paradigms** (i.e., semi-supervised, unsupervised representation, reinforcement learning) A major challenge of machine learning at scale is obtaining the pre-processed, labelled and large dataset [163]. To overcome this problem, algorithms such as semi-supervised and transfer learning are used. The former class of approaches increases the accuracy of the solution with less labelled data, and the latter by transferring the knowledge from the use-cases relevant to the current one.
- **Feature Engineering**: Selecting or engineering features that capture relevant information efficiently.
- **Dimensionality reduction**: Reducing data from a high-dimensional space to a lower-dimensional space to reduce computational complexity while retaining the (most) meaningful features. There exist diverse approaches, early ones like principal component analysis (PCA) or linear discriminant analysis (LDA) but also nonlinear and multi-dimensional ones [200].

14. OPEN QUESTIONS

In this last section, we present open questions and topics that were not covered in the initial version of this document. These sections may be included in subsequent versions of the document or remain as open questions. Obviously, this list is not exhaustive and is intended to encourage the submission of questions to the research departments of relevant universities or companies.

14.1 Does reusability make AI frugal?

Definition: In order to facilitate the widespread adoption of AI, it is imperative to explore approaches that can be readily implemented. A potential solution lies in the pre-training of AI models that can be either directly reused or rapidly customized to suit a variety of applications. Rather than developing a model from scratch, it would be more efficient and "expeditious" to assemble it from pre-existing components, analogous to the way in which we construct vehicles (cars, planes, etc.) by incorporating various parts.

Reusability[21] can improve the frugality of AI in several ways. Firstly, it promotes cost efficiency by reducing the need for extensive resources when training new models from scratch. In addition, it offers time savings by allowing developers to leverage existing solutions, which accelerates deployment. Furthermore, reusability helps optimize resources, minimizing both computational power and energy consumption. It also facilitates knowledge transfer, as reusable models can incorporate previously learned knowledge, improving performance without incurring additional training costs.

However, reusability may not always lead to frugality in AI. One concern is overfitting, where a model trained on a specific dataset may not generalize well to new data, potentially necessitating retraining. There are also maintenance costs associated with outdated or poorly designed reusable components, which can accumulate over time. Integration challenges may arise when reusing components from different projects, leading to compatibility issues that require additional resources to address. Moreover, the quality variability of reusable models can result in inefficiencies; not all models are of high quality, and using subpar options can increase long-term costs. Lastly, some applications might require significant customization of reused models, negating the initial cost savings.

Training reusable models is related to the challenge of creating models with strong generalization capabilities. A recent trend to enhance the generalizability of models, such as Large Language Models (LLMs), involves increasing the training compute and the size of the training dataset [35]. Although these approaches may seem fundamentally contrary to frugal principles, the upfront training cost can be amortized over multiple uses if these models are reused. Therefore, the trade-off between reusability and frugality should be considered when training such generalized models. Smaller but reusable pre-trained models, such as word2vec [160], should be encouraged.

This illustrates that while reusability has benefits, it can also lead to inefficiencies in certain contexts, opening up interesting research questions.

14.2 Does fine-tuning make AI frugal ?

Definition: "Fine-tuning" in AI refers to the process of taking a model pre-trained on a large dataset and making small adjustments to its parameters to adapt it to a specific, presumably smaller dataset [241; 227; 49]. The rationale is that the model benefits from the knowledge acquired during pre-training instead of starting from scratch, while still being tailored to the task of the smaller dataset.

Fine-tuning can contribute to making AI models more frugal in several ways: (i) reduced Training Time (fine-tuning a pre-trained model typically requires less time and computational resources compared to training a model from scratch); (ii) lower Data Requirements (fine-tuning often requires less data, as the model has already learned general features from the pre-training phase); (iii) efficiency in Resource Use (by leveraging existing knowledge, fine-tuned models can achieve good performance with fewer parameters, leading to lower memory and energy consumption).

Especially in terms of computational efficiency, several questions arise: (i) How does the training time for fine-tuning compare to training from scratch across various model architectures? What factors influence the efficiency of fine-tuning in terms of convergence speed and resource allocation? (ii) What strategies can be employed to further reduce data requirements during the fine-tuning process without sacrificing model performance? (iii) How does fine-tuning impact the memory and energy consumption of AI models in practical applications? What are the trade-offs between model size and performance when fine-tuning pre-trained models for specific tasks?

Note: Will most of the energy consumed by AI in 2025 be devoted to foundation models and fine-tuning even if they only cover part of the application of machine learning ?

[21] Maybe reusability is not limited to fine-tuning. In this case, a greater distinction would have to be made; a point we have not addressed in Sections 14.1 and 14.2.

Note 2 about sections 14.2 and 14.1: There are some overlapping ideas: (i) fine-tuning as part of a re-usability approach: in this case it can be understood under the prism of frugal AI because it means that one do not have to train models from scratch on large datasets (ii) fine-tuning as an obligatory step for LLMs: in this case it is rather 'anti-frugal' and this fine-tuning has more of a rebound effect.

14.3 Does making an AI sparse make it frugal?

Here, we use the following definition of a sparse AI model[22]:
Definition: A sparse AI model is a type of machine learning model that has a reduced number of model parameters or user parameters compared to its dense counterpart that can achieve the same task and for the same (or very close) performance.

The creation of an AI sparse model (e.g., using pruning methods, see Section 13.1) can result in a more frugal model in terms of resource usage. Sparse models generally require a reduced number of parameters and less computational power, which can result in decreased memory and energy consumption. However, it is important to note that the efficacy of sparsity depends on the specificity of the application and the model's ability to maintain performance despite reduced complexity. We may identify relevant questions and trade-offs regarding sparsity, particularly for those interested in deploying sparse models in real-world applications:

1. How does the sparsity level in AI models affect their performance across different sets of tasks? Are pruning methods task-dependent?
2. Are sparse models not only computationally more efficient but also more energy efficient than their dense counterparts? We emphasize this question because most of the engineering effort to deploy AI at scale is focused on dense models, and sparse models require different software architecture and hardware than their dense counterparts. Most notably, CPUs, instead of GPUs and TPUs, are known for being quite efficient on sparse computations [44].
3. Are sparse AI models more or less robust to adversarial attacks compared to their dense counterparts? In particular, gradient-based adversarial attacks are the most effective on dense models and modalities, such as images, in contrast to discrete modalities, such as textual data [231].
4. In which specific domains (e.g., natural language processing, computer vision) does sparsity provide the most significant benefits?

14.4 Should AI be resource-aware to be frugal?

Definition: "Resource-aware" refers to the ability of a system, application, or algorithm to recognize and efficiently utilize available resources, such as CPU, memory, bandwidth, and energy (for example some papers of the Lamarr Institute[23] as [62; 32] are on this topic). In the not-too-distant past, this approach to AI was known as 'ubiquitous learning' ([39; 232] see [link]).
Being resource-aware allows AI systems to (i) optimize resource utilization (efficiently allocate CPU, memory, and energy, etc.), (ii) adapt to constraints (adjust operations based on available resources, ...), (iii) fair usage of resources towards existing other applications in devices.
We may outline the following related questions: (i) What algorithms or techniques can be developed to enhance resource utilization in AI systems without compromising more or less performance? How do different AI architectures impact resource utilization efficiency, and what best practices can be established? (ii) How can AI models be designed to dynamically adjust their operations based on real-time resource availability? What are the implications of resource-aware adaptations on the accuracy and reliability of AI systems in various applications? (iii) What (new) metrics can be used to evaluate the sustainability of AI systems in terms of energy consumption and environmental impact?

14.5 How to explore effective strategies to circumvent the potential pitfalls of the rebound effect?

Definition: The AI rebound effect is defined as the phenomenon in which the efficiency or cost savings achieved through the utilisation of artificial intelligence result in an escalation in the consumption or utilisation of resources [19; 221; 222].
To illustrate this phenomenon, consider a scenario where AI is employed to enhance a process and reduce expenses. This may result in companies increasing their production or utilising additional resources, thereby negating the initial environmental or economic advantages. In summary, the rebound effect underscores the notion that enhancements in efficiency do not inherently ensure a decrease in overall impact. Interested readers can also consult section 3.3.

14.6 What social usages could bring to the frugal AI questioning?

In the context of increasing concerns about sustainability and resource efficiency, there is increasing concern about the use of frugal solutions and the promotion of low-tech technologies. These approaches advocate for simple, accessible, and often less costly methods that cater to local needs without necessitating complex infrastructures. By encouraging low-cost innovation and the use of local resources, these solutions promote greater social and economic inclusion. Furthermore, growing awareness of environmental issues is encouraging consumers and businesses to adopt solutions that minimize ecological impact, thereby reinforcing the acceptability of frugal and low-tech technologies as viable and responsible alternatives (related works [88]).

14.7 Frugal AI as a desirable side-effect of resource-constrained innovation?

Indeed, the implementation of frugal AI has the potential to result in the emergence of other priorities or requirements that may not have been the primary focus. To il-

[22] Even if all sparse models may not have a dense counterpart.

[23] https://lamarr-institute.org/research/resource-aware-ml/

lustrate this point, consider the context of the African market, where the adoption of frugal AI solutions is driven by specific challenges, including limited infrastructure, resource constraints, and diverse user needs. In such contexts, affordability, accessibility, and adaptability may take precedence over advanced features. Consequently, frugal AI can stimulate innovations tailored to local conditions, thereby fostering economic development and enhancing service delivery in sectors such as agriculture, healthcare, and education. Furthermore, it has the potential to encourage collaboration among local stakeholders, enhancing community engagement and ensuring that solutions are culturally relevant and sustainable.

Note: This question is also discussed in Section 5.

14.8 Will advancing learning theory result in more frugal AI models?

A specificity of the deployment of machine learning systems is that learning theory (i.e., theorems that give guarantees on the predictions made by AI systems upfront) lags behind the adoption of AI services across industries. This is not unprecedented in the history of technology; another such example is the steam machine, which drove the acceleration of the industrial revolution in the late 18th century, some 20 years before Carnot and other physicists gave a precise characterization of the thermodynamic laws in the early 19th century. Returning to machine learning, this raises the question of improved efficiency of AI systems driven by advances in learning theory.

As an illustrative example, there is a growing research effort toward understanding the complex interplay between memorization and generalization in machine learning: *generalization* refers to the ability to give accurate predictions on examples that have not been encountered during training, while *memorization* might be required in order to correctly classify rare instances [77], while also allowing for learning mislabelled examples which are arguably useless in order to solve the desired task [16; 84]. During training of a machine learning model, memorization takes the most of the compute time (thus, energy). This offers room for new strategies to mitigate unwanted memorization by focusing on better data curation.

Several research groups are examining this issue (see, for example, the [talk at Institute for Pure & Applied Mathematics (IPAM) of Gintarė Karolina Džiugaitė].

14.9 Can complex scalable systems be conceived as Frugal by design?

Energy production and consumption are closely related to environmental issues (air, water and thermal pollution, solid waste disposal, and climate change). However, the objective of the European Union to achieve carbon neutrality in 2050 is not achievable only by minimization of electrical energy [87]. To conceive frugal, scalable systems, we need to take into account the energy production/consumption aspects (devices, network, data centres) jointly with the eco-friendly device conception and the energy-efficient algorithms.

Two major research challenges linked with the energy consumption in AI from the perspective of scalable systems are (i) design of unified measures for energy consumption of various algorithms/hardware and (ii) evolution of unified measures sideways with new AI approaches and emerging technologies (edge-computing, quantum computing, generative AI, Agentic AI, or automatization/virtualization of future 6G networks).

Today, there is no unified tool that evaluates the energy consumption aspects for all use cases, usages, and data types, even if recent research efforts partially address this problem (i.e., training and inference evaluations of ML methods, [187], [209]). On the one hand, future research should focus on designing different types of frugal devices and systems from the hardware perspective (see Section 12 and its references). On the other side, research needs to design frugal methods that allow for the reuse of the existing resources whenever possible (i.e., multi-task training, transfer learning, or few-shot learning methods). The International standards committee for AI and the environment, among others, points out this duality between energy consumption and AI[24]: AI may consume a lot of energy (for example, deep learning, Generative AI or Agentic AI). However, it may also reduce the overall carbon footprint due to the reuse of a trained model in various fields.

Over the last decade, efficient methods at scale have been studied broadly (applications such as smart cities, connected vehicles, IoT). The energy efficiency of the algorithm has been shown to reduce the pollution and greenhouse gas emissions [87] by virtualisation, load balancing or consolidation. However, virtualization, softwarization and automatization of 5G and future 6G networks requires rethinking the design and usages of calculus (single data centers, hybrid or distributed approaches) in future research. Another research question is how to exploit the interconnection between the Power Grid that powers the networks, by considering the information on telecommunication network usages, that can be used to optimize the Power Grid [234], [4]. One example is how to use the energy metrics to predict the energy source availability, or how to use the prediction of energy source availability for optimal placement decisions.

The idea is also to think about complex systems that are designed from the outset to be frugal and scalable. To this end, they should incorporate a list of 'best practices'. These could include (but are not limited to): (i) minimalism: reducing unnecessary features and concentrating on essential functionality (ii) modularity: designing frugal components that can be easily modified or replaced without revising the whole system. The question is therefore to design a coherent and shared list of best practices and frugal components.

14.10 Will very large generative AIs (LLMs) and their uses one day become frugal?

The recent history of the Large Language Model (LLM) may give (instill) the impression that the larger the artificial intelligence system, the more useful it is. But this narrative obviously has a limit in terms of energy, material, infrastructure, network, ...[212; 25] The frugality of large-scale generative AI, (LSGenAI) is therefore an interesting question. This question is multifaceted since it can address: (i) the cost to pay to train an LSGenAI (ii) the cost to pay to use an LSGenAI (iii) the situation where LSGenAI are suitable[25] (iv) the sustainability of such AIS (v) all other ques-

[24] https://www.itu.int/dms_pub/itu-t/opb/env/T-ENV-ENV-2024-1-PDF-E.pdf

[25] For the fourth point, we refer the reader to the section 7

tions related to the cost of the infrastructure needed to 'run' them ... The purpose here is not how to avoid an "overshoot and collapse"[26] trajectory but rather how to create LSGenAI frugal by design? How to design them to incorporate some interesting facets (in a multi-criteria optimization [191]) by design as: (i) efficient architectures: utilizing streamlined model architectures (ii) data efficiency: training on smaller, high-quality datasets (iii) transfer Learning (iv) quantization (v) sparse models (vi) energy-efficient hardware...

14.11 Are there ways of thinking about the future of AI in a constrained environment?

Several scenarios for the ecological transition in 2050 emerge, including a frugal approach, a scenario focused on territorial cooperation, another focused on green technologies, and a last one, a repairing scenario. Each of these scenarios is expected to have different impacts on ecosystems. Consequently, examining the role of artificial intelligence in these different contexts may lead us to reassess our perspectives Surpassing planetary limits and their impact on the climate raises questions about the sustainability and future robustness of infrastructures and materials used in AI.

- *Which resource will be more critical for the future development of AI: electricity or rare metals? What are the physical limits of silicon chips, and how will this affect the future development of AI in a context of energy constraints?*
- *What strategies can be implemented to secure energy supply in the face of upcoming disruptions, particularly concerning AI?*
- *What tasks or jobs could AI replace in an energy-efficient manner in a world facing electricity constraints?*
- *What would tomorrow's business model be that could take account of these societal and environmental challenges?*
- *What would tomorrow's technologies be able to help in a constrained environment?*
- *How can we think about the impact of AI on society and the planet, by setting out governance principles and thinking about design to impact strategies?*

14.12 What could be frugal telecom network automation?

Network automation is seen as a key for operating operator's infrastructures, the Telco Management Forum (TMF) has defined 6 levels of automation each requiring more advanced architecture and technologies than the previous one. The trend to achieve level 4, is agentification and "LLM everywher" which comes at a significant environmental cost. Hence questionable when used for massive lower level machine to machine communication. While there is already a strong ongoing effort from an optimisation standpoint with protocols such as Agora [151] and the ability for agents to bypass LLMs with protocols such as MCP [104], some questions will of course remain when considering sustainable automation :

- What is the right level of automation for sustainable operations ? and how can we derive it from component performance ?
- What is the most efficient methodology to assess sustainability gains and impacts of automation ?
- Are there more frugal architectures that would still allow level 4 automation ?

14.13 Is semantic communication a means to frugal Agentic communications?

Context: Current multi-agent AI systems communicate mainly through conventional formats (JSON, UTF-8 encoded text), limiting their interactions to human-readable formats. However, these AI systems, particularly Large Language Models (LLMs), internally process information in rich semantic vector spaces. This creates an interesting paradox: while AI agents reason and process information in structured vector spaces, their communications are constrained to text-based exchanges.

The AI agent landscape is expected to expand significantly, from personal agents running on user's devices (smartphones, tablets) to enterprise-grade agents handling business operations, and service agents managing customer interactions. These agents will need to operate with increasing degrees of autonomy, making decisions and communicating with other agents to accomplish tasks without constant human supervision. The widespread deployment and autonomy of AI agents across various scales - from edge devices to cloud services - adds another dimension to the challenges of communication.

Hypothesis & Definition: It is hypothesised that future AI agent communications will evolve beyond text-based exchanges towards Semantic Communications, where agents directly transmit semantic representations (embeddings) through telecommunication networks. Semantic Communication involves the exchange of these structured vector representations that AI models use internally for processing information. This hypothesis is motivated by the nature of LLM processing, which occurs in structured vector spaces, and the limitations of current text-based communications in capturing the full semantic richness of AI representations. We therefore envision the emergence of new semantic "languages" shared between AI models, borrowing from those in-model representation spaces.

The adoption of semantic representations for inter-agent communications presents both opportunities and challenges for network frugality. On the one hand, these representations might enable more efficient and compact exchanges between AI agents, as semantic embeddings can encode complex meanings in structured ways, potentially reducing the number of exchanges needed for effective communication. On the other hand, the high-dimensionality of such representations (typically tens of KBytes per embedding) raises concerns about the network bandwidth required to support these communications, particularly in scenarios involving

of the present document.

[26] In the frugality context the idea of designing such frugal LSGenAI is not to try to solve the problem by producing more energy to consume more energy.

frequent exchanges between multiple autonomous agents at scale.

Open Research Questions on this topic (section):

- How can telecommunication networks efficiently support semantic communications between autonomous agents at scale?
- Can we develop specific encodings for semantic representations, similar to how audio and video codecs optimize media transmissions?
- What are the trade-offs between semantic fidelity and communication efficiency when compressing embeddings for inter-agent communication?
- What metrics can be developed to evaluate both the frugality and effectiveness of semantic communications?
- How can we ensure interoperability between different AI models and their semantic representations?

This list of thirteen questions, presented above, is obviously not exhaustive. If readers are interested in raising other ones, feel free to contact Nathalie Charbionnaud or Vincent Lemaire (firstname.name@orange.com).

15. REFERENCES

[1] Advanced Matrix Extensions. Advanced matrix extensions — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/wiki/Advanced_Matrix_Extensions`, 2024. [Online; accessed 2025-02-28].

[2] D. Aha, D. Kibler, and M. Albert. Instance-based learning algorithms, 1991.

[3] B. Ahat, A. C. Baktır, N. Aras, I. K. Altınel, A. Özgövde, and C. Ersoy. Optimal server and service deployment for multi-tier edge cloud computing. *Computer Networks*, 199:108393, 2021.

[4] S. Ahmed, T. M. Gondal, M. Adil, S. A. Malik, and R. Qureshi. A survey on communication technologies in smart grid. In *2019 IEEE PES GTD Grand International Conference and Exposition Asia (GTD Asia)*, pages 7–12, 2019.

[5] S. M. M. Ahsan, A. Dhungel, M. Chowdhury, M. S. Hasan, and T. Hoque. Hardware accelerators for artificial intelligence. *arXiv*, Nov. 2024.

[6] E. Ahvar, A.-C. Orgerie, and A. Lebre. Estimating energy consumption of cloud, fog and edge computing infrastructures. *IEEE Transactions on Sustainable Computing*, 7:277–288, 2022.

[7] S. Alam, C. Yakopcic, Q. Wu, M. Barnell, S. Khan, and T. M. Taha. Survey of deep learning accelerators for edge and emerging computing. *Electronics*, 13(15), 2024.

[8] G. Alavani, J. Desai, S. Saha, and S. Sarkar. Program analysis and machine learning–based approach to predict power consumption of cuda kernel. *ACM Trans. Model. Perform. Eval. Comput. Syst.*, 8(4), 2023.

[9] AMD. Amd versal^{TM} ai core series. `https://www.amd.com/en/products/adaptive-socs-and-fpgas/versal/ai-core-series.html`, 2025. [Online; accessed 2025-03-10].

[10] AMD. Amd vitis™ ai software. `https://www.amd.com/en/products/software/vitis-ai.html`, 2025. [Online; accessed 2025-03-10].

[11] A. Amirshahi, G. Ansaloni, and D. Atienza. Accelerator-driven data arrangement to minimize transformers run-time on multi-core architectures, 2023.

[12] Apache. Apache tvm. `https://tvm.apache.org/`, 2025. [Online; accessed 2025-03-18].

[13] Arm. Arm ethos-n hardware design. `https://developer.arm.com/Training/Arm%20Ethos-N%20Hardware%20Design`, 2025. [Online; accessed 2025-03-10].

[14] Arm. Cortex-m55. `https://developer.arm.com/processors/cortex-m55`, 2025. [Online; accessed 2025-03-10].

[15] ARM. Mali-g76. `https://developer.arm.com/Processors/Mali-G76`, 2025. [Online; accessed 2025-03-10].

[16] D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017.

[17] M. Arunachalam, V. Sanghavi, Y. A. Yao, Y. A. Zhou, L. A. Wang, Z. Wen, N. Ammbashankar, N. W. Wang, and F. Mohammad. Strategies for optimizing end-to-end artificial intelligence pipelines on intel xeon processors, 2022.

[18] E. Autret, N. Perry, M. Vautier, G. Busato, D. Charlet, M. Baccouche, G. Antipov, L. Charreire, V. Lemaire, P. Rust, L. Arga, T. Durand, U. Paila, and E. Abisset-Chavanne. IA et empreinte environnementale : Quelle consommation d'énergie pour quelles étapes ? Research report, 6, June 2022.

[19] I. M. Azevedo. Consumer end-use energy efficiency and rebound effects. *Annual Review of Environment and Resources*, 39(1):393–418, 2014.

[20] R. Balestriero, M. Ibrahim, V. Sobal, A. Morcos, S. Shekhar, T. Goldstein, F. Bordes, A. Bardes, G. Mialon, Y. Tian, A. Schwarzschild, A. G. Wilson, J. Geiping, Q. Garrido, P. Fernandez, A. Bar, H. Pirsiavash, Y. LeCun, and M. Goldblum. A cookbook of self-supervised learning, 2023.

[21] K. E. Bassey, A. R. Juliet, and A. O. Stephen. Ai-enhanced lifecycle assessment of renewable energy systems. *Engineering Science & Technology Journal*, 2024.

[22] O. Bause, P. P. Bernardo, and O. Bringmann. A configurable and efficient memory hierarchy for neural network hardware accelerator, 2024.

[23] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, and U. M. . Greedy layer-wise training of deep networks. *Advances in Neural Information Processing Systems*, 19, 01 2007.

[24] Berkeley Lab. 2024 United States Data Center Energy Usage Report, 2024. https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report.pdf.

[25] E. Bhardwaj, R. Alexander, and C. Becker. Limits to ai growth: The ecological and social consequences of scaling, 2025.

[26] G. Bhattacharya. From dnns to gans: Review of efficient hardware architectures for deep learning, 2021.

[27] W. L. Bircher and L. K. John. Complete system power estimation using processor performance events. *IEEE Transactions on Computers*, 61(4):563–577, 2012.

[28] T. Bolukbasi, J. Wang, O. Dekel, and V. Saligrama. Adaptive neural networks for efficient inference, 2017.

[29] M. Boullé. Khiops: outil d'apprentissage supervisé automatique pour la fouille de grandes bases de données multi-tables. *Revue des Nouvelles Technologies de l'Information*, Extraction et Gestion des Connaissances, RNTI-E-30:505–510, 2016. www.khiops.org.

[30] R. Bourgeot. Sommet de l'ia de bletchley park : Concertation mondiale ou lobbying chic?. *IRIS*, November 2023. https://www.iris-france.org/179597-sommet-de-lia-de-bletchley-park-concertation-mondiale-ou-lobbying-chic/.

[31] G. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1976.

[32] R. Braun, Tanya; Möller. Lessons from resource-aware machine learning for healthcare: An interview with katharina morik. *KI - Künstliche Intelligenz*, 38:243–248, March 2024.

[33] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[34] H. Brendan McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*, 54, 2017.

[35] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[36] T. Bu, W. Fang, J. Ding, P. L. Dai, Z. Yu, and T. Huang. Optimal Ann-Snn Conversion for High-Accuracy and Ultra-Low-Latency Spiking Neural Networks. *ICLR 2022 - 10th International Conference on Learning Representations*, 2022.

[37] J. G. Burns, J. Foucaud, and F. Mery. Costs of memory: Lessons from 'mini' brains. *Proceedings of the Royal Society B: Biological Sciences*, 278(1707):923–929, 2011.

[38] Candt Solution. What is asic? application specific integrated circuits. https://www.candtsolution.com/news_events-detail/what-is-asic-application-specific-integrated-circuits/, 2024. [Online; accessed 2025-02-28].

[39] L. A. Cárdenas-Robledo and A. Peña-Ayala. Ubiquitous learning: A systematic review. *Telematics and Informatics*, 35(5):1097–1132, 2018.

[40] R. Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.

[41] F. Chance. Lessons from *a* dragon fly's brain: Evolution built a small, fast, efficient neural network in a dragonfly. why not copy it for missile defense? *IEEE Spectrum*, 58(8):28–33, 2021.

[42] F. S. Chance. Interception from a Dragonfly Neural Network Model. *ACM International Conference Proceeding Series*, 2020.

[43] L. Chauveau. La sncf sur la voie du solaire, February 2025. https://www.sciencesetavenir.fr/high-tech/transports/la-sncf-veut-alimenter-ses-trains-avec-l-energie-solaire_183875.

[44] B. Chen, T. Medini, J. Farwell, C. Tai, A. Shrivastava, et al. Slide: In defense of smart algorithms over hardware acceleration for large-scale deep learning systems. *Proceedings of Machine Learning and Systems*, 2:291–306, 2020.

[45] H. Cheng, M. Zhang, and J. Q. Shi. A Survey on Deep Neural Network Pruning: Taxonomy, Comparison, Analysis, and Recommendations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(8):1–30, 2024.

[46] L. Chittka and J. Niven. Are Bigger Brains Better? *Current Biology*, 19(21):R995–R1008, 2009.

[47] K. T. Chitty-Venkata, M. Emani, V. Vishwanath, and A. K. Somani. Neural Architecture Search Benchmarks: Insights and Survey. *IEEE Access*, 11(March):25217–25236, 2023.

[48] A. Christopher. The future of data science jobs: Will 2030 mark their end?, 2024. https://medium.com/dataseries/the-future-of-data-science-jobs-will-2030-mark-their-end-d01b1a52ce4a.

[49] K. W. Church, Z. Chen, and Y. Ma. Emerging trends: A gentle introduction to fine-tuning. *Natural Language Engineering*, 27(6):763–778, 2021.

[50] G. Cloud. Cloud tensor processing units (tpus). https://cloud.google.com/tpu?hl=pl, 2025. [Online; accessed 2025-03-10].

[51] B. Collins. Nvidia ceo predicts the death of coding — jensen huang says ai will do the work, so kids don't need to learn, 2024. https://www.techradar.com/pro/nvidia-ceo-predicts-the-death-of-coding-jensen-huang-says-ai-will-do-the-work-so-kids-dont-need-to-learn.

[52] M. . Company. The state of ai in early 2024: Gen ai adoption spikes and starts to generate value, 2024. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai.

[53] S. J. Cook, T. A. Jarrell, C. A. Brittin, Y. Wang, A. E. Bloniarz, M. A. Yakovlev, K. C. Nguyen, L. T.-H. Tang, E. A. Bayer, J. S. Duerr, et al. Whole-animal connectomes of both caenorhabditis elegans sexes. *Nature*, 571(7763):63–71, 2019.

[54] A. S. Corporation. achronix. https://www.achronix.com/, 2025. [Online; accessed 2025-03-10].

[55] H. Corporation. Ascend computing. https://e.huawei.com/pl/products/computing/ascend, 2025. [Online; accessed 2025-03-10].

[56] I. H. Corporation. Ai & compute. https://www.imaginationtech.com/products/ai/, 2025. [Online; accessed 2025-03-10].

[57] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

[58] F. Cortney. A Survey on Network Quantization Techniques for Deep Neural Network Compression, 2024.

[59] B. Courty, V. Schmidt, S. Luccioni, Goyal-Kamal, MarionCoutarel, B. Feld, J. Lecourt, LiamConnell, A. Saboni, Inimaz, supatomic, M. Léval, L. Blanche, A. Cruveiller, ouminasara, F. Zhao, A. Joshi, A. Bogroff, H. de Lavoreille, N. Laskaris, E. Abati, D. Blank, Z. Wang, A. Catovic, M. Alencon, Michał Stęchły, C. Bauer, L. O. N. de Araújo, JPW, and MinervaBooks. mlco2/codecarbon: v2.4.1, May 2024. https://doi.org/10.5281/zenodo.11171501.

[60] M. Crépel and D. Cardon. Robots vs algorithms: Prophétie et critique dans la représentation médiatique des controverses de l'ia. *Réseaux*, 232-233(2):129–167, 2022.

[61] R. Dattakumar and R. Jagadeesh. A review of literature on benchmarking. *Benchmarking: An International Journal*, 10(3):176–209, 2003.

[62] A. Dave, F. Frustaci, F. Spagnolo, M. Yayla, J.-J. Chen, and H. Amrouch. Hw/sw codesign for approximation-aware binary neural networks. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 13(1):33–47, 2023.

[63] I. N. de l'Audovisuel. Source de données ina. url(https://data.ina.fr/cles-lecture/words).

[64] D. De Silva and D. Alahakoon. An artificial intelligence life cycle: From conception to production. *Patterns*, 3(6):100489, 2022.

[65] C. B. Delahunt and J. N. Kutz. Putting a bug in ML: The moth olfactory network learns to read MNIST. *Neural Networks*, 118:54–64, 2019.

[66] H. Dempsey. World's largest transformer maker warns of supply crunch. https://www.ft.com/content/a0fa2e61-b684-42b7-bd12-6b9d7c28285c.

[67] J. Derise. Will the data industry continue to consolidate?, 2024. https://thedatascore.substack.com/p/will-the-data-industry-continue-to.

[68] C. Desroches, M. Chauvin, L. Ladan, C. Vateau, S. Gosset, and P. Cordier. Exploring the sustainable scaling of ai dilemma: A projective study of corporations' ai environmental impacts, 01 2025.

[69] I.-C. E. d'ingénieurs. Etude ipsos - intelligence artificielle : quels sont les usages des français ?, February 2025. https://www.ipsos.com/fr-fr/intelligence-artificielle-quels-sont-les-usages-des-francais.

[70] S. Dorkenwald, A. Matsliah, A. R. Sterling, P. Schlegel, S.-C. Yu, C. E. McKellar, A. Lin, M. Costa, K. Eichler, Y. Yin, et al. Neuronal wiring diagram of an adult brain. *Nature*, 634(8032):124–138, 2024.

[71] Z. Durante, Q. Huang, N. Wake, R. Gong, J. S. Park, B. Sarkar, R. Taori, Y. Noda, D. Terzopoulos, Y. Choi, K. Ikeuchi, H. Vo, L. Fei-Fei, and J. Gao. Agent ai: Surveying the horizons of multimodal interaction, 2024.

[72] A. EC2. Amazon ec2 f2. https://aws.amazon.com/ec2/instance-types/f2/, 2025. [Online; accessed 2025-03-10].

[73] G. Elimian. Chatgpt costs $700,000 to run daily, openai may go bankrupt in 2024, 2023. https://technext24.com/2023/08/14/chatgpt-costs-700000-daily-openai/.

[74] T. Elsken, J. H. Metzen, and F. Hutter. Neural architecture search: A survey. *Journal of Machine Learning Research*, 20:1–21, 2019.

[75] J. Falgas and P. Robert. Présenter l'ia comme une évidence, c'est empêcher de réfléchir le numérique. *The Conversation*, February 2025. http://theconversation.com/presenter-lia-comme-une-evidence-cest-empecher-de-reflechir-le-numerique-211766.

[76] J. Fang, Y. Shen, Y. Wang, and L. Chen. Optimizing dnn computation graph using graph substitutions. *Proc. VLDB Endow.*, 13(12):2734–2746, 2020.

[77] V. Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 954–959, 2020.

[78] I. for Orange. Sociovisions 2024 - ifop pour orange, 2024.

[79] V. for SII. L'intelligence artificielle et les français - viavoice pour sii, February 2024. https://sii-roup.com/sites/default/files/document/SII_Sondage_IA_2024.pdf.

[80] I. for Talan. Baromètre 2024 "les français et les ia génératives' vague 2 – ifop pour talan, may 2024. https://www.ifop.com/wp-content/uploads/2024/07/120717-Rapport-reduit.pdf.

[81] R. France. La production de l'électricité, June 2022. https://assets.rte-france.com/prod/public/2022-06/FE2050%20_Rapport%20complet_4.pdf.

[82] K. Fukushima. Visual feature extraction by a multilayered network of analog threshold elements. *IEEE Transactions on Systems Science and Cybernetics*, 5(4):322–333, 1969.

[83] M. Garnelo and M. Shanahan. Reconciling deep learning with symbolic artificial intelligence: representing objects and relations. *Current Opinion in Behavioral Sciences*, 29:17–23, 2019. Artificial Intelligence.

[84] T. George, P. Nodet, A. Bondu, and V. Lemaire. Mislabeled examples detection viewed as probing machine learning models: concepts, survey and extensive benchmark. *Transactions on Machine Learning Research*, 2024.

[85] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer. A Survey of Quantization Methods for Efficient Neural Network Inference. *Low-Power Computer Vision*, pages 291–326, 2022.

[86] GlobalData. Globaldata, generative ai market report, 2024. https://www.globaldata.com.

[87] W. E. Gnibga, A. Blavette, and A.-C. Orgerie. Renewable energy in data centers: The dilemma of electrical grid dependency and autonomy costs. *IEEE Transactions on Sustainable Computing*, 9(3):315–328, 2024.

[88] K. Govindan. How artificial intelligence drives sustainable frugal innovation: A multitheoretical perspective. *IEEE Transactions on Engineering Management*, 71:638–655, 2022.

[89] S. R. Group. Hyperscale capacity set to triple by 2030, 2025. https://www.srgresearch.com/articles/hyperscale-data-center-capacity-to-triple-by-2030-driven-by-generative-ai.

[90] G. Guidi, F. Dominici, J. Gilmour, K. Butler, E. Bell, S. Delaney, and F. J. Bargagli-Stoffi. Environmental burden of united states data centers in the artificial intelligence era, 2024.

[91] H. Guillaud. Comprendre ce que l'ia sait faire et ce qu'elle ne peut pas faire. https://danslesalgorithmes.net/2024/10/10/comprendre-ce-que-lia-sait-faire-et-ce-quelle-ne-peut-pas-faire/, 2024. [Online; accessed 2025-03-].

[92] O. Gupta and R. Raskar. Distributed learning of deep neural network over multiple agents, 2018.

[93] A. Hanna and E. M. Bender. Ai causes real harm. let's focus on that over the end-of-humanity hype. https://www.scientificamerican.com/article/we-need-to-focus-on-ais-real-harms-not-imaginary-existential-risks/, 2024. [Online; accessed 2025-03-].

[94] A. Hanna and E. M. Bender. Ai causes real harm. let's focus on that over the end-of-humanity hype. *Scientific American*, February 2024. https://www.scientificamerican.com/article/we-need-to-focus-on-ais-real-harms-not-imaginary-existential-risks/.

[95] S. Hanyu. The combination of metal oxides as oxide layers for rram and artificial intelligence, 2023.

[96] H. Hassan, S. Barakat, and Q. Sarhan. Survey on serverless computing. *Journal of Cloud Computing volume 10*, 2021.

[97] Y. He and L. Xiao. Structured Pruning for Deep Convolutional Neural Networks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):2900–2919, 2024.

[98] D. W. Heaven. What is ai? https://www.technologyreview.com/2024/07/10/1094475/what-is-artificial-intelligence-ai-definitive-guide/, 2024. [Online; accessed 2025-03-24].

[99] W. D. Heaven. What is ai? *MIT Technology Review*, July 2024.

[100] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554, 2006.

[101] S. C. H. Hoi, D. Sahoo, J. Lu, and P. Zhao. Online learning: A comprehensive survey, 2018.

[102] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey. Meta-learning in neural networks: A survey, 2020.

[103] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey. Meta-learning in neural networks: A survey, 2020.

[104] X. Hou, Y. Zhao, S. Wang, and H. Wang. Model context protocol (mcp): Landscape, security threats, and future research directions, 2025.

[105] K. Hu. Chatgpt sets record for fastest-growing user base, 2023. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

[106] B. K. Hulse, H. Haberkern, R. Franconville, D. Turner-Evans, S. Y. Takemura, T. Wolff, M. Noorman, M. Dreher, C. Dan, R. Parekh, A. M. Hermundstad, G. M. Rubin, and V. J. V. A connectome of the drosophila central complex reveals network motifs suitable for flexible navigation and context-dependent action selection. *eLife*, 10:1–180, 2021.

[107] E. J. Husom, A. Goknil, L. K. Shar, and S. Sen. The price of prompting: Profiling energy use in large language models inference, 2024. https://www.arxiv.org/abs/2407.16893.

[108] S. Ibrahim, H. Hazimeh, and R. Mazumder. Flexible modeling and multitask learning using differentiable tree ensembles, 2022.

[109] Intel. Altera® fpga and soc fpga. https://www.intel.com/content/www/us/en/products/details/fpga.html, 2025. [Online; accessed 2025-03-10].

[110] Intel. Intel vision accelerator design with intel® movidius™ vision processing unit (vpu). https://www.intel.com/content/www/us/en/developer/topic-technology/edge-5g/hardware/vision-accelerator-movidius-vpu.html, 2025. [Online; accessed 2025-03-10].

[111] Intel. Welcome to intel® gaudi® v1.20 documentation. https://docs.habana.ai/en/latest/, 2025. [Online; accessed 2025-03-10].

[112] International Energy Agency. Efficiency improvement of AI related computer chips, 2008-2023, October 2024. https://www.iea.org/data-and-statistics/charts/efficiency-improvement-of-ai-related-computer-chips-2008-2023.

[113] International Energy Agency. World Energy Outlook 2024, 2024. https://www.iea.org/reports/world-energy-outlook-2024/executive-summary?language=en.

[114] L. Jacob, F. Bach, and J.-P. Vert. Clustered multi-task learning: A convex formulation, 2008.

[115] M. Jay, V. Ostapenco, L. Lefevre, D. Trystram, A.-C. Orgerie, and B. Fichel. An experimental comparison of software-based power meters: focus on cpu and gpu. In *2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*, pages 106–118, 2023.

[116] Z. Jia, J. Thomas, T. Warszawski, M. Gao, M. Zaharia, and A. Aiken. Optimizing dnn computation with relaxed graph substitutions. In A. Talwalkar, V. Smith, and M. Zaharia, editors, *Proceedings of Machine Learning and Systems*, volume 1, pages 27–39, 2019.

[117] L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey, 1996.

[118] Y. Kang, J. Hauswald, C. Gao, A. Rovinski, T. Mudge, J. Mars, and L. Tang. Neurosurgeon: Collaborative intelligence between the cloud and mobile edge. *SIGPLAN Not.*, 52(4):615–629, 2017.

[119] A. Kansal, F. Zhao, J. Liu, N. Kothari, and A. A. Bhattacharya. Virtual machine power metering and provisioning. In *Proceedings of the 1st ACM Symposium on Cloud Computing*, SoCC '10, page 39–50, New York, NY, USA, 2010. Association for Computing Machinery.

[120] J. Kelly. Goldman sachs predicts 300 million jobs will be lost or degraded by artificial intelligence, 2023. https://www.forbes.com/sites/jackkelly/2023/03/31/goldman-sachs-predicts-300-million-jobs-will-be-lost-or-degraded-by-artificial-intelligence/.

[121] Khadas. Vim3. https://www.khadas.com/vim3, 2025. [Online; accessed 2025-03-10].

[122] A. Kiachian. Nvidia launches generative ai microservices for developers to create and deploy generative ai copilots across nvidia cuda gpu installed base, 2024. https://nvidianews.nvidia.com/news/generative-ai-microservices-for-developers.

[123] K. Kim and M.-j. Park. Present and future, challenges of high bandwith memory (hbm). In *2024 IEEE International Memory Workshop (IMW)*, pages 1–4, 2024.

[124] M. Kinder. Hollywood writers went on strike to protect their livelihoods from generative ai. their remarkable victory matters for all workers., 2024. https://www.brookings.edu/articles/hollywood-writers-went-on-strike-to-protect-their-livelihoods-from-generative-ai-their-remarkable-victory-matters-for-all-workers/.

[125] A. Klimczak, M. Wenka, M. Ganzha, M. Paprzycki, and J. Mańdziuk. Towards frugal artificial intelligence: Exploring neural network pruning and binarization. In S. M. Thampi, J. Mukhopadhyay, M. Paprzycki, and K.-C. Li, editors, *International Symposium on Intelligent Informatics*, pages 13–27, Singapore, 2023. Springer Nature Singapore.

[126] W. Klöpffer and B. Grahl. *Life cycle assessment (LCA): a guide to best practice*. John Wiley & Sons, 2014.

[127] J. Konečný, H. B. McMahan, D. Ramage, and P. Richtaric. Federated optimization: Distributed machine learning for on-device intelligence. *http://arxiv.org/abs/1610.02527*, pages 1–38, 2016.

[128] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon. Federated learning: Strategies for improving communication efficiency. *http://arxiv.org/abs/1610.05492*, pages 1–10, 2017.

[129] A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres. Quantifying the carbon emissions of machine learning. In *NeurIPS 2019 Workshop on Tackling Climate Change with Machine Learning*, 2019.

[130] G. Lagani, F. Falchi, C. Gennaro, and G. Amato. Spiking Neural Networks and Bio-Inspired Supervised Deep Learning: A Survey. *Asian Journal of Research in Computer Science*, pages 1–31, 2023.

[131] Y. LeCun, Y. Bengio, and G. E. Hinton. Deep learning, 2015.

[132] C. Lee. China is on course for a prolonged recession, 2025. https://www.aspistrategist.org.au/china-is-on-course-for-a-prolonged-recession/.

[133] J. K. Lee, L. Mukhanov, A. S. Molahosseini, U. Minhas, Y. Hua, J. Martinez Del Rincon, K. Dichev, C. H. Hong, and H. Vandierendonck. Resource-Efficient Convolutional Networks: A Survey on Model-, Arithmetic-, and Implementation-Level Techniques. *ACM Computing Surveys*, 55(13 s), 2023.

[134] T. Legrand. Deepseek vs chatgpt: The comprehensive 2025 comparison shaking up the ai industry, 2025. `https://www.digidop.com/blog/deepseek-vs-chatgpt`.

[135] V. Lemaire, F. Clérot, N. Voisine, C. Hue, F. Fessant, R. Trinquart, and F. Olmos Marchan. The data mining process : a (not so) short introduction, 2017. `https://www.researchgate.net/publication/313528093_The_Data_Mining_Process_a_not_so_short_introduction`.

[136] F. Li, J. Lindsey, E. C. Marin, N. Otto, M. Dreher, G. Dempsey, I. Stark, A. S. Bates, M. W. Pleijzier, P. Schlegel, A. Nern, S. Takemura, N. Eckstein, T. Yang, A. Francis, A. Braun, R. Parekh, M. Costa, L. Scheffer, Y. Aso, G. S. Jefferis, L. F. Abbott, A. Litwin-Kumar, S. Waddell, and G. M. Rubin. The connectome of the adult drosophila mushroom body provides insights into function. *eLife*, 9:1–217, 2020.

[137] H. H. Li. Ai models for edge computing: Hardware-aware optimizations for efficiency. In *2024 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1–1. IEEE, 2024.

[138] Y. Li, Y. Lei, and X. Yang. Spikeformer: A Novel Architecture for Training High-Performance Low-Latency Spiking Neural Network. *arXiv.org*, 2022.

[139] Y. Li, S. Wang, Y. Zhao, S. Wang, W. Zhang, Y. He, N. Lin, B. Cui, X. Chen, S. Zhang, H. Jiang, P. Lin, X. Zhang, X. Qi, Z. Wang, X. Xu, D. Shang, Q. Liu, K.-T. Cheng, and M. Liu. Pruning random resistive memory for optimizing analogue ai, 2023.

[140] Z. Li, C. Yan, X. Zhang, G. Gharibi, Z. Yin, X. Jiang, and B. A. Malin. Split learning for distributed collaborative training of deep learning models in health informatics, 2023.

[141] B.-S. Liang. Design of asic accelerators for ai applications. *IET conference proceedings.*, 2024(19):147–154, Jan. 2025.

[142] J. Liang. Design and optimization of hardware accelerators for convolutional neural networks. *Science and technology of engineering, chemistry and environmental protection*, 1(10), Dec. 2024.

[143] Z. Lin, M. Courbariaux, R. Memisevic, and Y. Bengio. Neural networks with few multiplications. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, pages 1–9, 2016.

[144] J. Z. Lingjiao Chen, Matei Zaharia. Frugalgpt: How to use large language models while reducing cost and improving performance, 2024. `https://arxiv.org/abs/2305.05176`.

[145] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In W. W. Cohen and H. Hirsh, editors, *Machine Learning, Proceedings of the Eleventh International Conference, Rutgers University, New Brunswick, NJ, USA, July 10-13, 1994*, pages 157–163. Morgan Kaufmann, 1994.

[146] A. S. Luccioni, S. Viguier, and A.-L. Ligozat. Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. *Journal of Machine Learning Research*, 24(253):1–15, 2023.

[147] make.org. Consultation citoyenne : What are your ideas for shaping ai to serve the public good – make.org pour sciences po, ai & society institute (ens-psl), the future society, cnum, December 2024.

[148] C. Malone and C. Belady. Metrics to characterize data center & it equipment energy use. In *Proceedings of the Digital Power Forum*, 2006.

[149] E. M. Marianne TORDEUX BITKER. Pour une intelligence artificielle au service de l'intérêt général, 2025. `https://www.lecese.fr/travaux-publies/pour-une-intelligence-artificielle-au-service-de-linteret-general`.

[150] G. C. Marinó, A. Petrini, D. Malchiodi, and M. Frasca. Deep neural networks compression: A comparative survey and choice recommendations. *Neurocomputing*, 520:152–170, 2023.

[151] S. Marro, E. L. Malfa, J. Wright, G. Li, N. Shadbolt, M. Wooldridge, and P. Torr. A scalable communication protocol for networks of large language models, 2024.

[152] N.-E. Mbengue. Une mesure de l'empreinte environnementale des modèles d'ia pour une utilisation plus frugale, 2024. `https://management-datascience.org/articles/31291/`.

[153] N. E. Mbengue. Étude comparative de l'empreinte carbone de modèles de machine learning appliqués au traitement automatique de la langue (tal). Master's thesis, TELECOM Nancy, 2024.

[154] W. S. Mcculloch and W. Pitts. A logical calculus nervous activity. *Bulletin of Mathematical Biology*, 52(l):99–115, 1990.

[155] Mckinsey. Global Survey on AI, 1,363 participants at all levels of the organization, February 2024.

[156] MediaTek. Mediatek dimensity 5g. `https://www.mediatek.com/products/smartphones/dimensity-5g`, 2025. [Online; accessed 2025-03-10].

[157] V. Mehlin, S. Schacht, and C. Lanquillon. Towards energy-efficient Deep Learning: An overview of energy-efficient approaches along the Deep Learning Lifecycle. *arXiv*, 2023.

[158] G. Menghani. Efficient Deep Learning: A Survey on Making Deep Learning Models Smaller, Faster, and Better. *ACM Computing Surveys*, pages 1–36, 2023.

[159] L. Messeri and M. J. Crockett. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58, 2024. https://doi.org/10.1038/s41586-024-07146-0.

[160] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[161] R. Mishra and H. Gupta. Transforming Large-Size to Lightweight Deep Neural Networks for IoT Applications. *ACM Computing Surveys*, 55(11), 2023.

[162] A. Moslemi, A. Briskina, Z. Dang, and J. Li. Machine Learning with Applications A survey on knowledge distillation : Recent advancements. *Machine Learning with Applications*, 18(November), 2024.

[163] P. Nodet, V. Lemaire, A. Bondu, A. Cornuéjols, and A. Ouorou. From weakly supervised learning to biquality learning: an introduction. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–10, 2021.

[164] J. D. Nunes, M. Carvalho, D. Carneiro, and J. S. Cardoso. Spiking Neural Networks: A Survey. *IEEE Access*, 10:60738–60764, 2022.

[165] Nvidia. Jetson modules. https://developer.nvidia.com/embedded/jetson-modules, 2025. [Online; accessed 2025-03-10].

[166] Nvidia. Nvidia tensorrt. https://developer.nvidia.com/tensorrt, 2025. [Online; accessed 2025-03-18].

[167] ONNX. Open neural network exchange. https://onnx.ai/, 2025. [Online; accessed 2025-03-18].

[168] OpenVINO™Toolkit. Openvino. https://github.com/openvinotoolkit/openvino, 2025. [Online; accessed 2025-03-18].

[169] OpenXLA. Xla. https://openxla.org/xla, 2025. [Online; accessed 2025-03-18].

[170] V. Ostapenco, L. Lefèvre, A.-C. Orgerie, and B. Fichel. Modeling, evaluating and orchestrating heterogeneous environmental leverages for large scale data centers management. *International Journal of High Performance Computing Applications, SAGE*, 37:328–350, 2023.

[171] X. Ou, Z. Chen, C. Zhu, and Y. Liu. Low Rank Optimization for Efficient Deep Learning: Making a Balance Between Compact Architecture And Fast Training. *Journal of Systems Engineering and Electronics*, 35(3):509–531, 2023.

[172] P. D. H. P., N. Gillis, and X. Siebert. A survey on deep matrix factorizations. *Computer Science Review*, 42, 2021.

[173] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[174] D. Paul, G. Namperumal, and A. Selvaraj. Cloud-native ai/ml pipelines: Best practices for continuous integration, deployment, and monitoring in enterprise applications. *Journal of Artificial Intelligence Research*, 2(1):176–230, 2022.

[175] R.-D. Pinzon-Morales and Y. Hirata. Cerebellar-inspired bi-hemispheric neural network for adaptive control of an unstable robot. In *2013 ISSNIP Biosignals and Biorobotics Conference: Biosignals and Robotics for Better and Safer Living (BRC)*, pages 1–4. IEEE, 2013.

[176] Planète Energies. Les modes de production de l'électricité, 2023. https://www.planete-energies.com/fr/media/article/production-delectricite-ses-emissions-co2.

[177] L. O. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin. Catboost: unbiased boosting with categorical features. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *NeurIPS*, pages 6639–6649, 2018.

[178] T. Pröottel. WSTS World Semiconductors Trade Statistics (11-2023), Gartner, IBS and Tech Insights forecast, January 2024. https://www.wsts.org/.

[179] Qualcomm. Qualcomm adreno gpu. https://www.qualcomm.com/products/technology/processors/adreno, 2025. [Online; accessed 2025-03-10].

[180] M. Rabin. Le béton est une source majeure du réchauffement climatique, 2023. https://reporterre.net/Le-beton-est-une-source-majeure-du-rechauffement-climatique.

[181] H. Rapp and M. P. Nawrot. A spiking neural program for sensorimotor control during foraging in flying insects. *Proceedings of the National Academy of Sciences of the United States of America*, 117(45):28412–28421, 2020.

[182] S. Ren. How much water does ai consume? the public deserves to know, 2023. https://oecd.ai/en/wonk/how-much-water-does-ai-consume.

[183] J. Requeima, J. Gordon, J. Bronskill, S. Nowozin, and R. E. Turner. Fast and flexible multi-task classification using conditional neural adaptive processes. *Advances in neural information processing systems*, 32, 2019.

[184] A. Rios and P. Nava. Hardware for quantized mixed-precision deep neural networks. In *2022 IEEE 15th Dallas Circuit And System Conference (DCAS)*, pages 1–5. IEEE, 2022.

[185] S. Rivoire, P. Ranganathan, and C. E. Kozyrakis. A comparison of high-level full-system power models. In *Power-Aware Computer Systems*, 2008.

[186] Rock-Chips. Rk3399pro. https://www.rock-chips.com/a/en/products/RK33_Series/2018/0130/874.html, 2025. [Online; accessed 2025-03-10].

[187] C. Rodriguez, L. Degioanni, L. Kameni, R. Vidal, and G. Neglia. Evaluating the energy consumption of machine learning: Systematic literature review and experiments, 2024.

[188] B. Rokh, A. Azarpeyvand, and A. Khanteymoori. A Comprehensive Survey on Model Quantization for Deep Neural Networks in Image Classification. *ACM Transactions on Intelligent Systems and Technology*, 14(6), 2023.

[189] P. Ruvolo and E. Eaton. ELLA: An efficient lifelong learning algorithm, 17–19 Jun 2013.

[190] K. Ryan, Z. Lu, and I. A. Meinertzhagen. The cns connectome of a tadpole larva of ciona intestinalis (l.) highlights sidedness in the brain of a chordate sibling. *Elife*, 5:e16962, 2016.

[191] H. M. Salkin and C. A. De Kluyver. The knapsack problem: a survey. *Naval Research Logistics Quarterly*, 22(1):127–144, 1975.

[192] D. Saul. Are we suddenly close to a recession? here's what the data actually shows, 2025. https://www.forbes.com/sites/dereksaul/2025/03/08/are-we-suddenly-close-to-a-recession-heres-what-the-data-actually-shows/.

[193] P. Schlegel, A. S. Bates, T. Stürner, S. R. Jagannathan, N. Drummond, J. Hsu, L. S. Capdevila, A. Javier, E. C. Marin, A. Barth-Maron, I. F. Tamimi, F. Li, G. M. Rubin, S. M. Plaza, M. Costa, and G. S. Jefferis. Information flow, cell types and stereotypy in a full olfactory connectome. *eLife*, 10:1–47, 2021.

[194] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni. Green ai, 2019.

[195] C. J. Shallue, J. Lee, J. Antognini, J. Sohl-Dickstein, R. Frostig, and G. E. Dahl. Measuring the effects of data parallelism on neural network training, 2019.

[196] A. Shawahna, S. M. Sait, and A. El-Maleh. Fpga-based accelerators of deep learning networks for learning and classification: A review. *IEEE Access*, 7:7823–7859, 2019.

[197] Y. Shen, S. Dasgupta, and S. Navlakha. Algorithmic insights on continual learning from fruit flies, 2021.

[198] Z. Skidmore. Are nuclear power and SMRs the solution to data center energy woes?, November 2024. https://www.datacenterdynamics.com/en/analysis/nuclear-power-smr-us/.

[199] D. Snider and R. Liang. Operator fusion in xla: Analysis and evaluation, 2023.

[200] C. O. S. Sorzano, J. Vargas, and A. P. Montano. A survey of dimensionality reduction techniques. *arXiv preprint arXiv:1403.2877*, 2014.

[201] R. S. Sutton and A. G. Barto. Reinforcement learning: An introduction, 2018.

[202] C. Séramour. IA générative : Microsoft relance la centrale nucléaire de Three Mile Island pour alimenter ses data centers, september 2024. https://www.usine-digitale.fr/article/ia-generative-microsoft-relance-la-centrale-nucleaire-de-three-mile-island-pour-alimenter-ses-data-centers.N2219114.

[203] T. Tahmaseb. Preparing for the 2030s depression, 2024. https://blog.itreconomics.com/blog/preparing-for-the-2030s-depression.

[204] S. Takemura, A. Bharioke, Z. Lu, and al. A visual motion detection circuit suggested by Drosophila connectomics. *Nature*, 500(7461):175–181, 2013.

[205] A. Tang, P. Quan, L. Niu, and Y. Shi. A survey of sparse regularization based compression methods. *Procedia Computer Science*, 199(2021):703–709, 2021.

[206] Y. Tang, Y. Wang, J. Guo, Z. Tu, K. Han, H. Hu, and D. Tao. A Survey on Transformer Compression, 2024.

[207] S. Teerapittayanon, B. McDanel, and H. T. Kung. Branchynet: Fast inference via early exiting from deep neural networks, 2017.

[208] S. Thrun and L. Pratt. *Learning to learn: introduction and overview*, page 3–17. Kluwer Academic Publishers, USA, 1998.

[209] C. E. Tripp, J. Perr-Sauer, J. Gafur, A. Nag, A. Purkayastha, S. Zisman, and E. A. Bensen. Measuring the energy consumption and efficiency of deep neural networks: An empirical analysis and design recommendations, 2024.

[210] S. Vadera and S. Ameen. Methods for Pruning Deep Neural Networks. *IEEE Access*, 10:63280–63300, 2022.

[211] M. Vaithianathan. Memory hierarchy optimization strategies for high- performance computing architectures. *International Journal of Emerging Trends & Technology in Computer Science*, pages 1–24, 01 2025.

[212] G. Varoquaux, A. S. Luccioni, and M. Whittaker. Hype, sustainability, and the price of the bigger-is-better paradigm in ai, 2025.

[213] C. Verasztó, S. Jasek, M. Gühmann, R. Shahidi, N. Ueda, J. D. Beard, S. Mendes, K. Heinz, L. A. Bezares-Calderón, E. Williams, et al. Whole-animal connectome and cell-type complement of the three-segmented platynereis dumerilii larva. *BioRxiv*, pages 2020–08, 2020.

[214] A. Vicente-Sola, D. L. Manna, P. Kirkland, G. Di Caterina, and T. Bihl. Spiking Neural Networks for event-based action recognition: A new task to understand their advantage, 2022.

[215] A. Vijayan and S. Diwakar. A cerebellum inspired spiking neural network as a multi-model for pattern classification and robotic trajectory prediction. *Frontiers in Neuroscience*, 16:909146, 2022.

[216] B. Villalonga, D. Lyakh, S. Boixo, H. Neven, T. S. Humble, R. Biswas, E. G. Rieffel, A. Ho, and S. Mandr. Establishing the quantum supremacy frontier with a 281 Pflop/s simulation. *Quantum Science and Technology*, 5(3):1–14, 2020.

[217] Z. Wang, P. Wang, K. Liu, P. Wang, Y. Fu, C.-T. Lu, C. C. Aggarwal, J. Pei, and Y. Zhou. A comprehensive survey on data augmentation, 2024.

[218] C. Warzel. Ai has become a technology of faith. *The Atlantic*, July 2024. `https://www.theatlantic.com/technology/archive/2024/07/thrive-ai-health-huffington-altman-faith/678984/`.

[219] K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big data*, 3(1):9, 2016.

[220] C. White, M. Safari, R. Sukthanker, B. Ru, T. Elsken, A. Zela, D. Dey, and F. Hutter. Neural Architecture Search: Insights from 1000 Papers, 2023.

[221] M. Willenbacher, T. Hornauer, and V. Wohlgemuth. Rebound effects in methods of artificial intelligence. In *Environmental Informatics*, pages 73–85. Springer, 2021.

[222] M. Willenbacher, T. Hornauer, and V. Wohlgemuth. A short overview of rebound effects in methods of artificial intelligence. *Int. J. Environ. Sci. Nat. Res*, 2021.

[223] M. Winding, B. D. Pedigo, C. L. Barnes, H. G. Patsolic, Y. Park, T. Kazimiers, A. Fushiki, I. V. Andrade, A. Khandelwal, J. Valdes-Aleman, et al. The connectome of an insect brain. *Science*, 379(6636):eadd9330, 2023.

[224] M. E. Wright. Ai 2020: The global state of intelligent enterprise. `https://www.intelligentautomation.network/artificial-intelligence/whitepapers/i2020`. Accessed: 2025-02-10.

[225] B. Xia, Q. Lu, L. Zhu, and Z. Xing. An ai system evaluation framework for advancing ai safety: Terminology, taxonomy, lifecycle mapping. In *Proceedings of the 1st ACM International Conference on AI-Powered Software*, New York, NY, USA, 2024. Association for Computing Machinery.

[226] Y. Xiao, C. Gao, J. Jin, W. Sun, B. Wang, Y. Bao, C. Liu, W. Huang, H. Zeng, and Y. Yu. Recent Progress in Neuromorphic Computing from Memristive Devices to Neuromorphic Chips. *Advanced Devices and Instrumentation*, 5, 2024.

[227] Y. Xin, J. Yang, S. Luo, H. Zhou, J. Du, X. Liu, Y. Fan, Q. Li, and Y. Du. Parameter-efficient fine-tuning for pre-trained vision models: A survey, 2025.

[228] Y. Xing, G. Di Caterina, and J. Soraghan. A New Spiking Convolutional Recurrent Neural Network (SCRNN) With Applications to Event-Based Hand Gesture Recognition. *Frontiers in Neuroscience*, 14(November), 2020.

[229] C. Xu and J. McAuley. A Survey on Model Compression and Acceleration for Pretrained Language Models. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023*, volume 37, pages 10566–10575, 2023.

[230] G. Xu, W. Huang, and W. Jia. A comprehensive survey on recent model compression and acceleration approaches for deep neural networks and transformers. *Available at SSRN 4893335*, 2024.

[231] H. Xu, Y. Ma, H.-C. Liu, D. Deb, H. Liu, J.-L. Tang, and A. K. Jain. Adversarial attacks and defenses in images, graphs and text: A review. *International journal of automation and computing*, 17:151–178, 2020.

[232] S. Yahya, E. Ahmad, and K. Abd Jalil. The definition and characteristics of ubiquitous learning: A discussion. *International Journal of Education and Development using ICT*, 6(1), 2010.

[233] W. S. Yamamoto and T. B. Achacoso. Scaling up the nervous system of caenorhabditis elegans: is one ape equal to 33 million worms? *Computers and biomedical research*, 25(3):279–291, 1992.

[234] Y. Yan, Y. Qian, H. Sharif, and D. Tipper. A survey on smart grid communication infrastructures: Motivations, requirements and challenges. *Communications Surveys & Tutorials, IEEE*, 15:5–20, 01 2013.

[235] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, 2016.

[236] A. M. Zador. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications*, 10(1), 2019.

[237] K. Zaman, A. Marchisio, M. A. Hanif, and M. Shafique. A Survey on Quantum Machine Learning: Current Trends, Challenges, Opportunities, and the Road Ahead, 2023.

[238] Z. Zeya. Qingdao port automated terminal sets record-breaking performance for the 12th time, 2025. `http://en.sasac.gov.cn/2025/01/16/c_18725.htm`.

[239] D. Zhang and J. F. Nunamaker. Powering e-learning in the new millennium: An overview of e-learning and enabling technology. *Information Systems Frontiers*, 5:207–218, 2003.

[240] T. Zhang, Z. Li, Y. Chen, K.-Y. Lam, and J. Zhao. Edge-cloud cooperation for dnn inference via reinforcement learning and supervised learning. In *2022 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing ; Communications (GreenCom) and IEEE Cyber, Physical ; Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)*, page 77–84. IEEE, Aug. 2022.

[241] H. Zheng, L. Shen, A. Tang, Y. Luo, H. Hu, B. Du, Y. Wen, and D. Tao. Learning from models beyond fine-tuning. *Nature Machine Intelligence*, 7(1):6–17, Jan. 2025.

[242] H. Zhou, M. Li, N. Wang, G. Min, and J. Wu. Accelerating deep learning inference via model parallelism and partial computation offloading. *IEEE Transactions on Parallel and Distributed Systems*, 34(2):475–488, 2023.

[243] B. Zhu. L'ia embarquée ouvre la voie à la prochaine révolution de l'ia, 2024. `https://www.allnews.ch/partenaires/content/l%E2%80%99ia-embarqu%C3%A9e-ouvre-la-voie-%C3%A0-la-prochaine-r%C3%A9volution-de-l%E2%80%99ia`.

[244] X. Zhu, J. Li, Y. Liu, C. Ma, and W. Wang. A Survey on Model Compression for Large Language Models. In *Transactions of the Association for Computational Linguistics*, volume 12, pages 1556–1577, 2024.

[245] I. Zliobaitė, M. Budka, and F. Stahl. Towards cost-sensitive adaptation: When is it worth updating your predictive model? *Neurocomputing*, 150:240–249, 2015. Bioinspired and knowledge based techniques and applications The Vitality of Pattern Recognition and Image Analysis Data Stream Classification and Big Data Analytics.