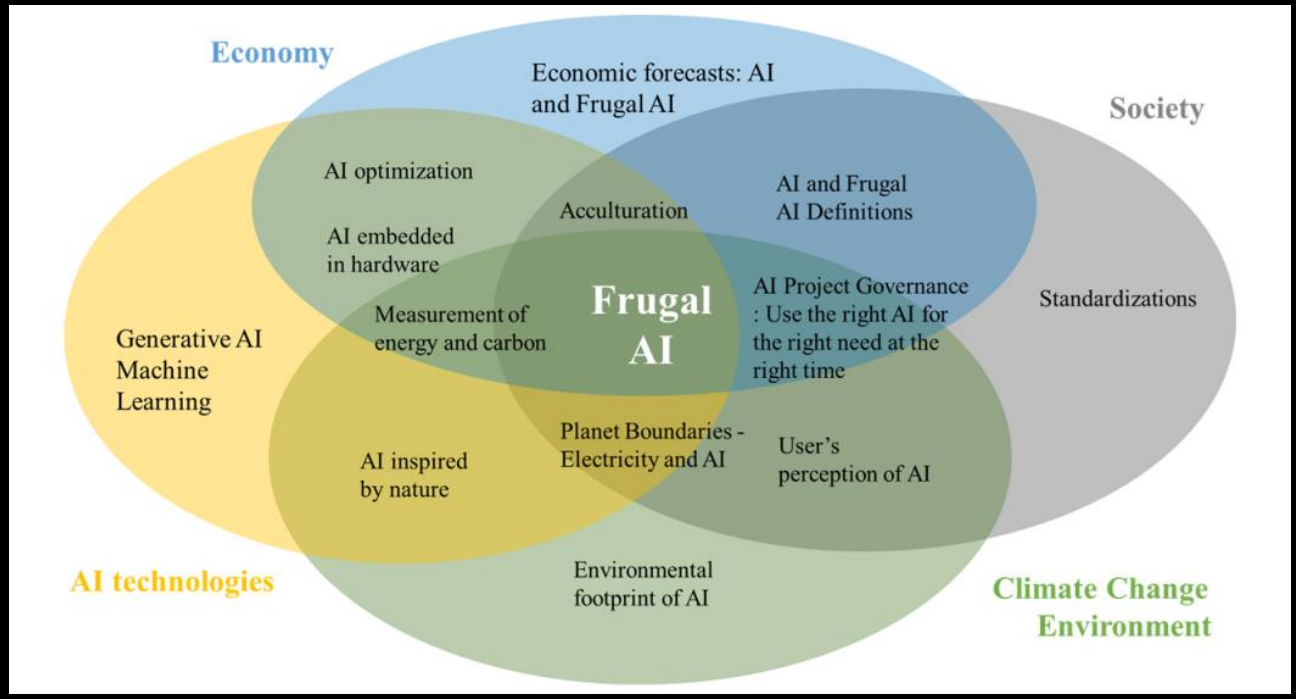
Frugal AI

Vincent LEMAIRE

Nicolas Voisine



orange™

PFIA 2026

Une vision parcellaire

Grand Merci à : Nathalie Charbonniaud, Paweł Piotrowski, Krzysztof Sapiejewski, Tamara Tosic, Frédéric Walch, Romain Carbou…

Orange Restricted

Plutôt une discussion autour d'une table plutôt qu'un tutoriel descendant

Orange Restricted

Première partie plutôt factuelle

Orange Restricted

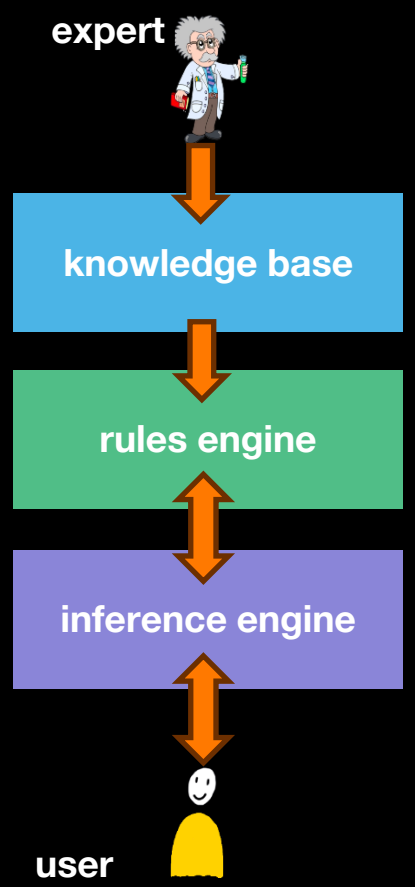
Few words about AI

Vocabulary, history, typology, trends

Orange Restricted

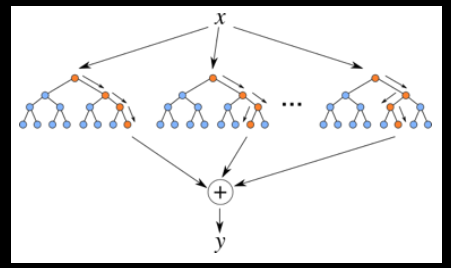
What exactly does “AI” cover?

expert system 19xx

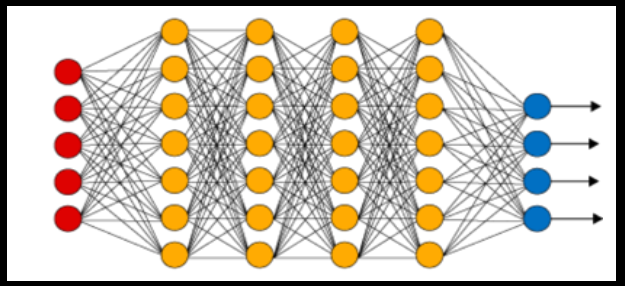


machine learning 19xx

- supervised learning (classification)
- unsupervised learning (clustering)
- reinforcement learning



deep learning 201x



transformers 2017

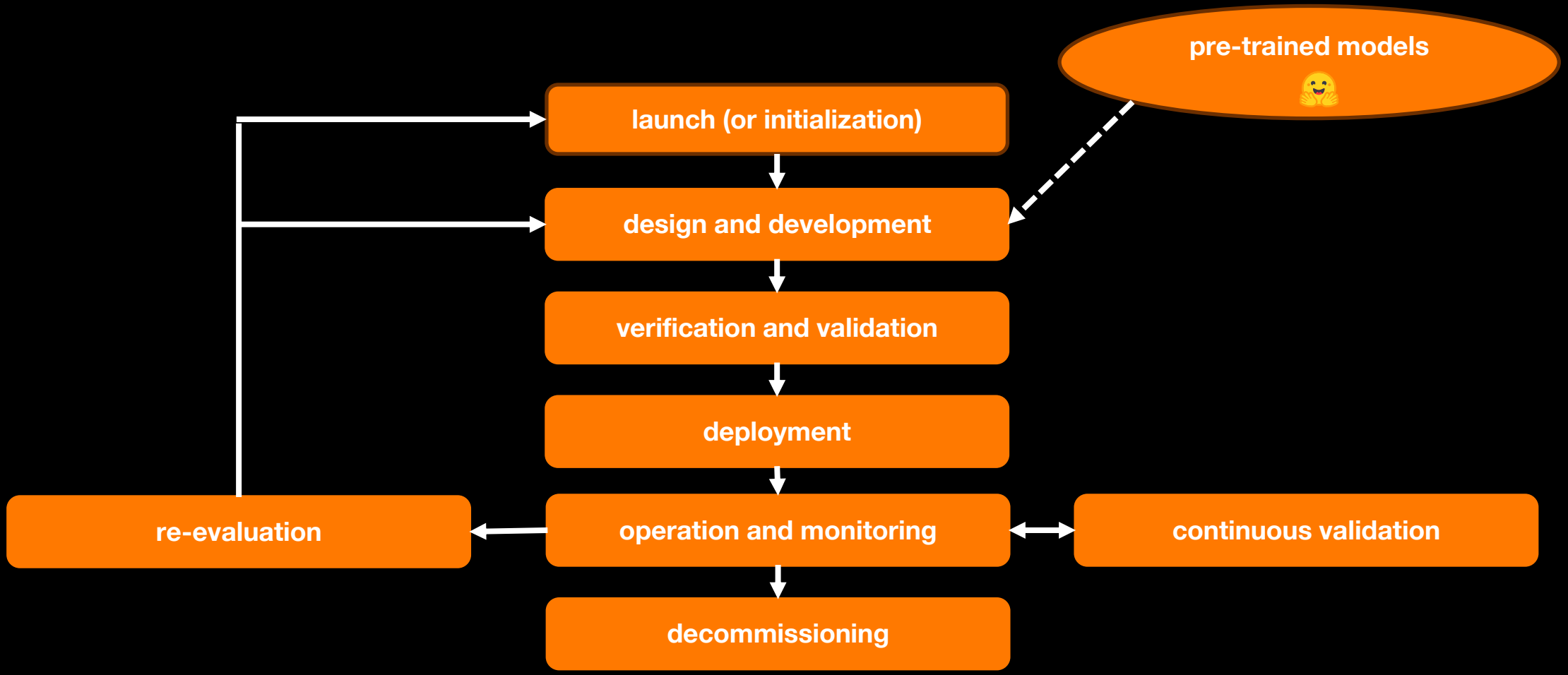
LLM, Generative AI (“GenAI”) 2022

Gemini

General AI (“AGI”) 20??

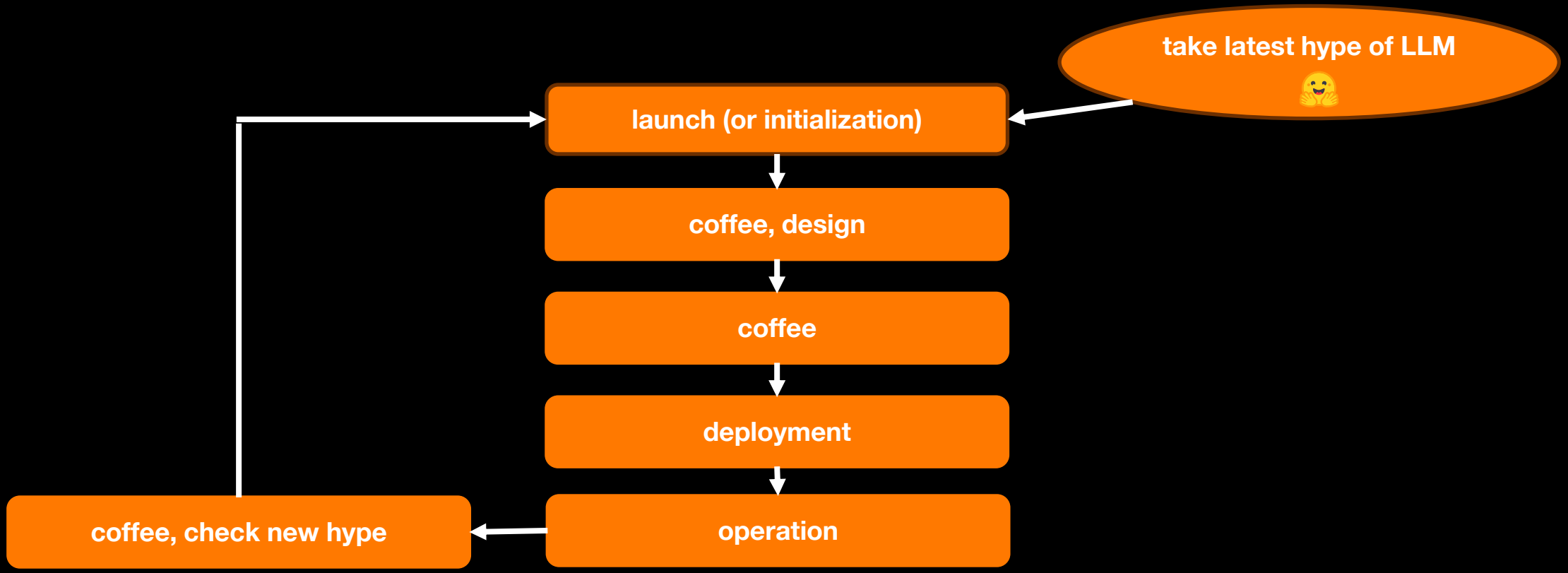
Orange Restricted

Life cycle of an AI system (ML / DL)



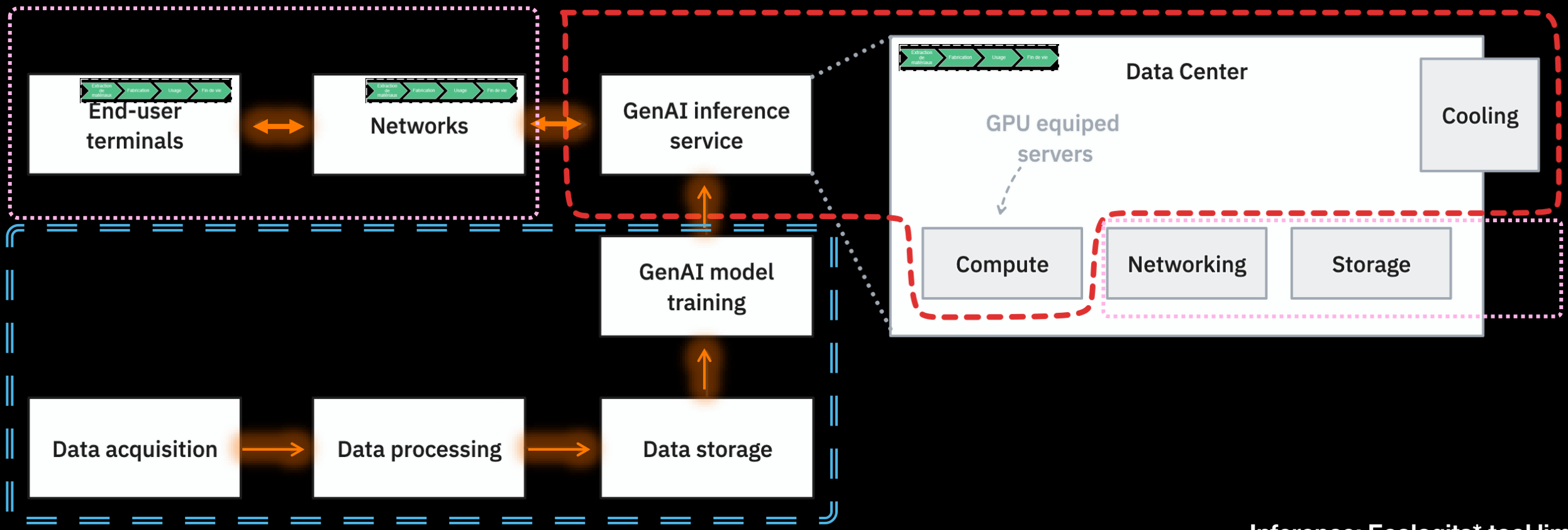
Orange Restricted

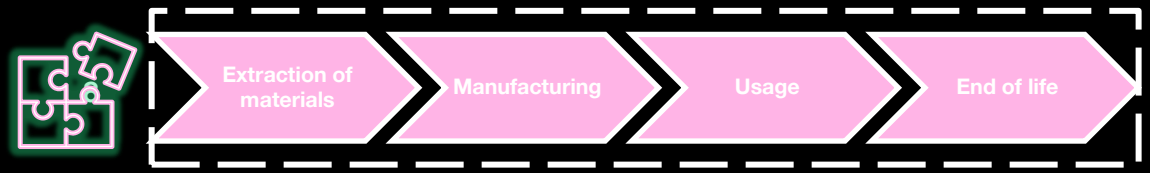
Life cycle of some GenAI-based projects



Orange Restricted

Carbon footprint e2e of an AI service

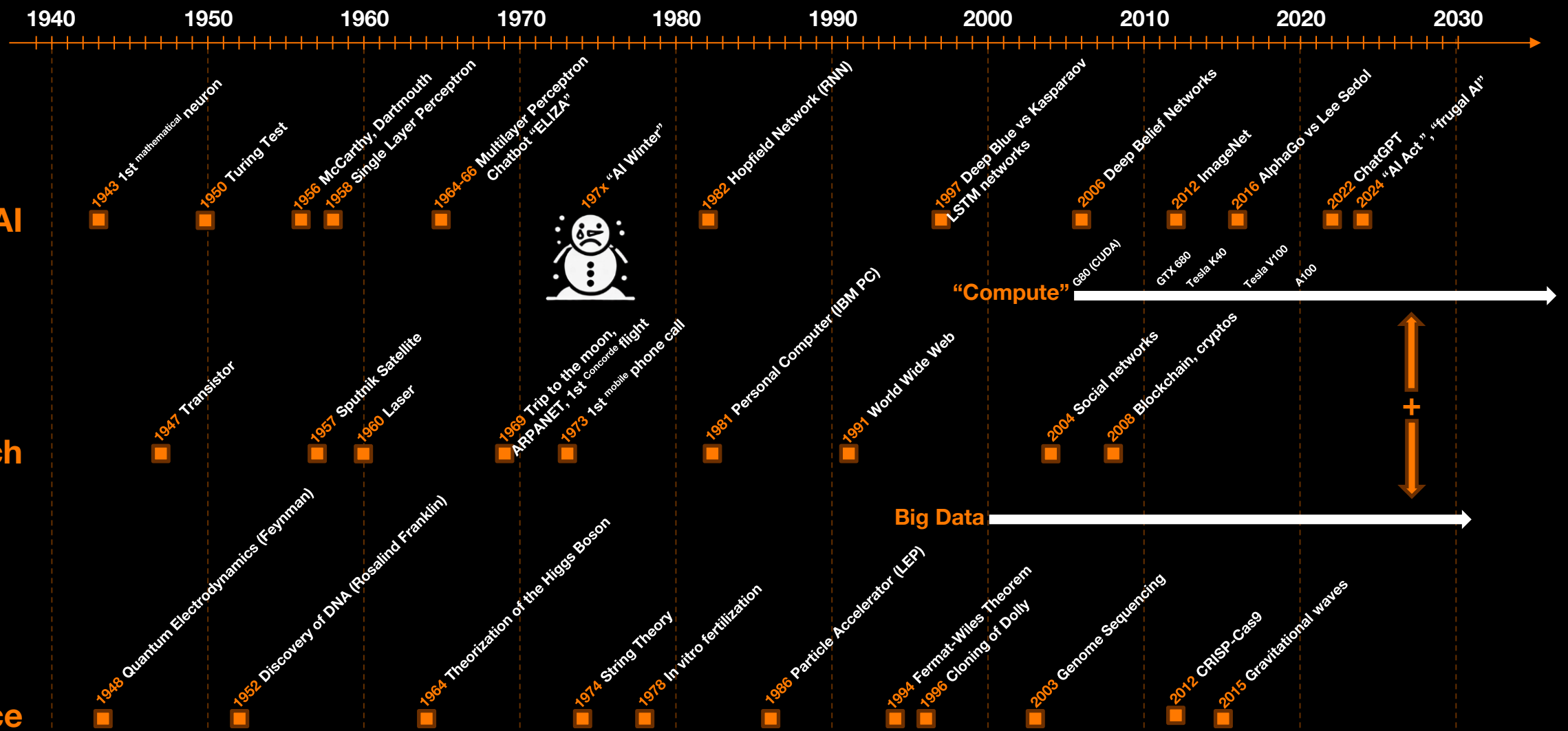




--- Inference: Ecologits* tool limit
.... Add to measure digital service footprint
== Model training (and re-training)
--- Lifecycle stages of each digital tier

*Source : EcoLogits Calculator - a Hugging Face Space by genai-impact

Timeline(s)



Orange Restricted

The current situation

Orange Restricted

Material context

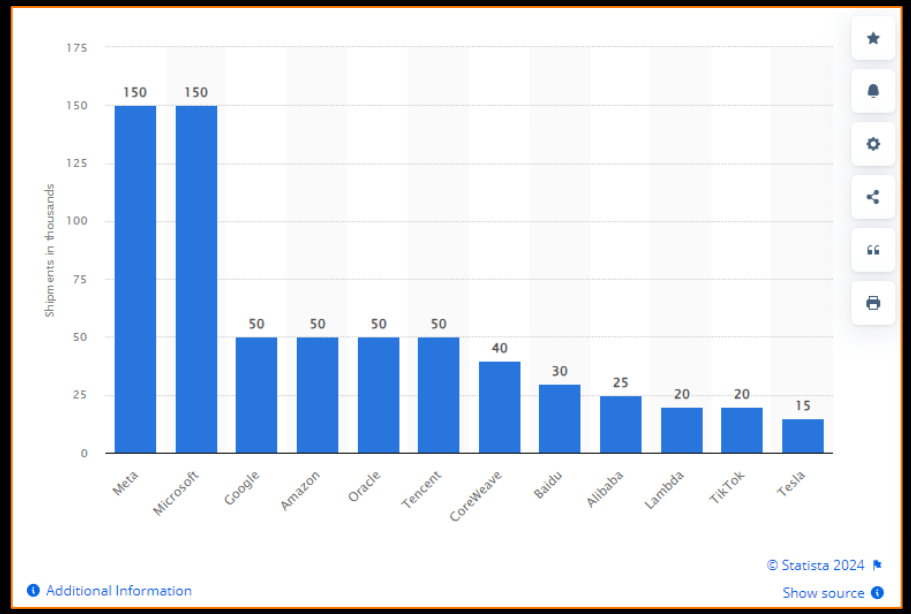
The revolution of Deep Learning and then of generative AI is a waste of resources.

Manufacturing of 1 dedicated AI server 3,700 kg $CO2_{eq}$ = round trip Paris-NY / person

ChatGPT's (assumed) infrastructure ~ 3,500 servers ~ 28,000 GPUs

NVIDIA

A (very) small number of players are monopolizing the production of high-end GPUs



Orange Restricted

Environmental context: water

IT and AI in particular are not just about consuming resources and energy.

- Data centers need water to prevent overheating
- New Meta data center planned in Spain:
  - 102 hectares of hangars in a protected nature reserve
  - 600 million liters of water / year
  - of which, 200 million are taken from the drinking water network of the neighbouring town (Talavera de la Reina)

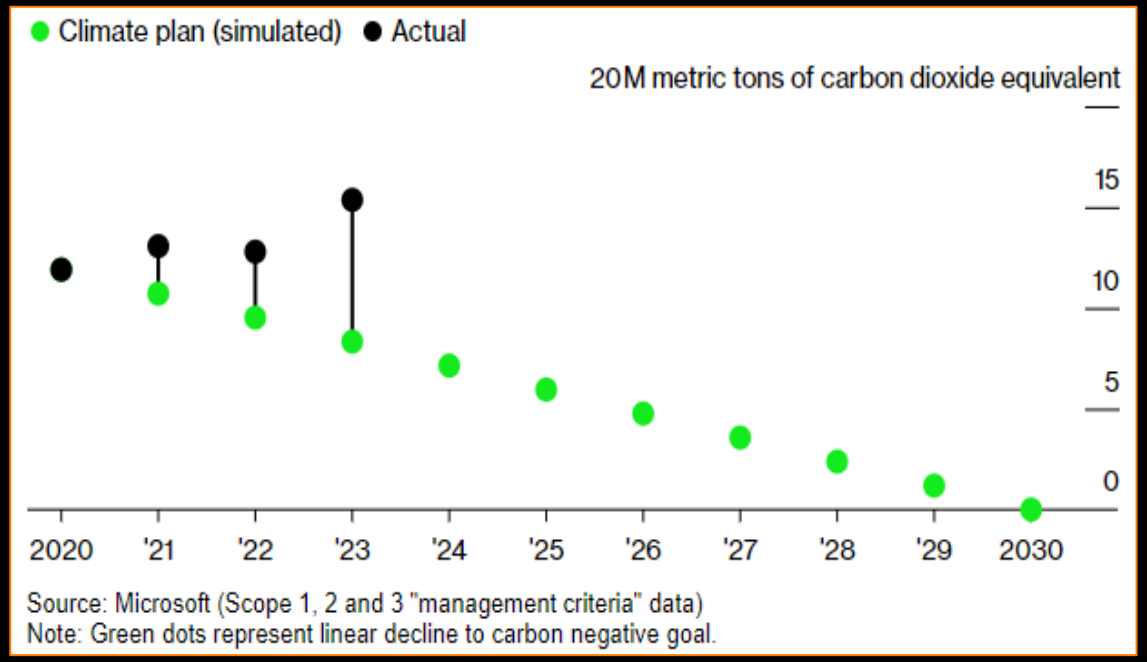
- Uruguayan protesters opposed to a Google data center project, against the backdrop of a major water crisis
  - Paso Severino reservoir level at about 2% of capacity, public water is no longer drinkable
  - 7 million liters of water / day

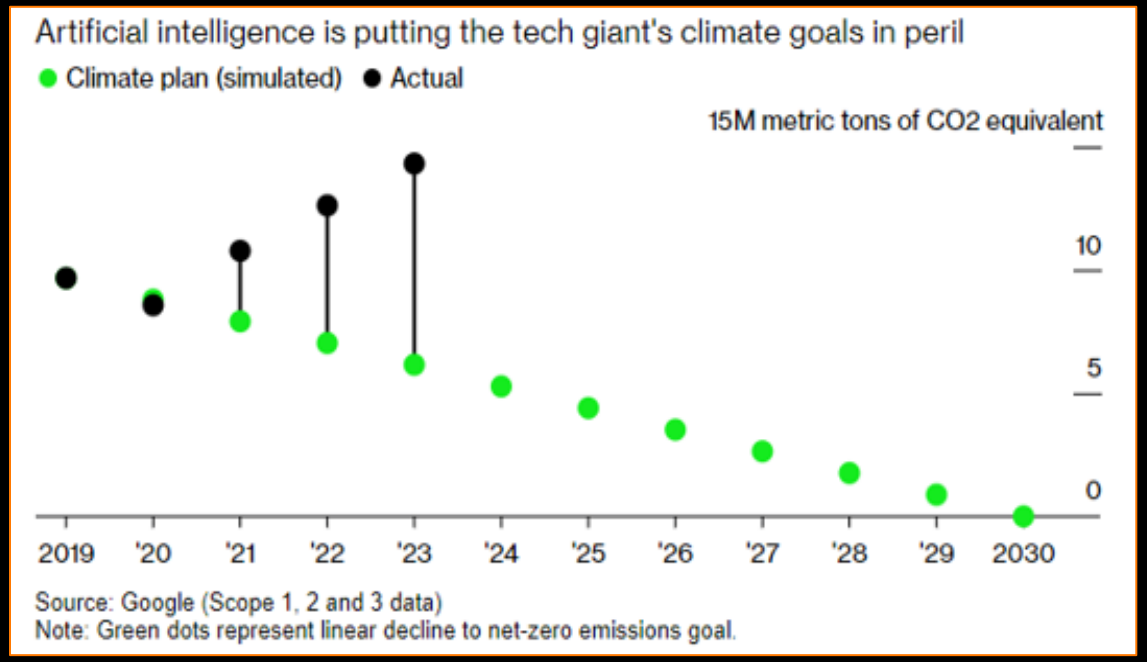
- 2023 : Generative AI generated 2,600 tonnes of “W3E” (*)
- 2030 : x 1,000?

Orange Restricted

(*) waste electrical and electronic equipment

Carbon footprint: what about GAFAM?





Microsoft
+ 34% CO_2eq
2020 → 2023

Google
+ 48% CO_2eq
2019 → 2023

Orange Restricted

Quiz: ChatGPT 4o-mini Footprint

If 1% of Google searches were done with ChatGPT 4o-mini (~400 tokens), what would it amount to in terms of carbon impact?

A. 50 times around the Earth by car

B. 600 times around the Earth by car

5 295,9 T CO_2 eq (608 times actually ☺)

C. 200 times around the Earth by car

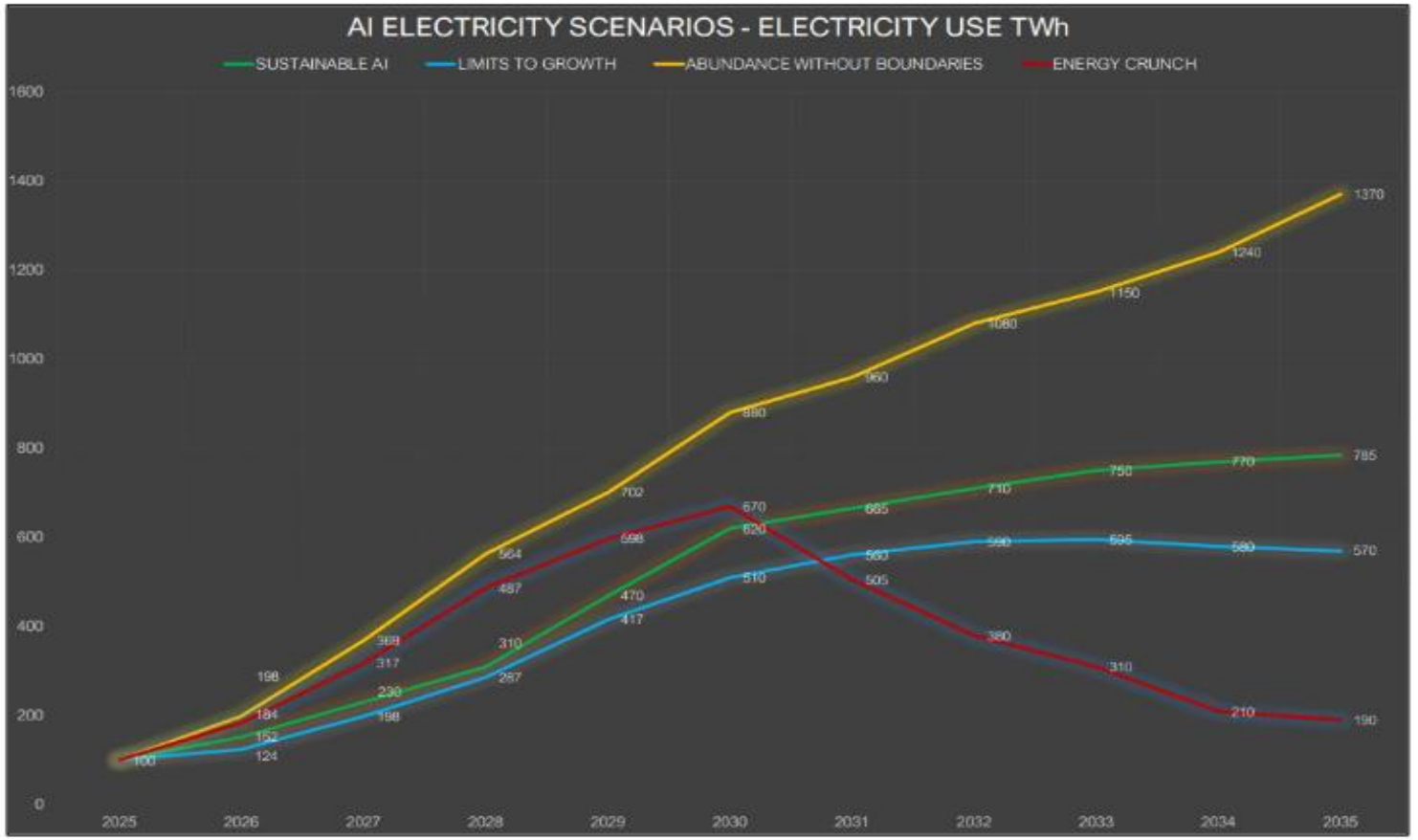
ChatGPT = 800 million active users / month!
1 month of use = 20 model trainings

Orange Restricted

Energy consumption prediction:

Intelligence artificielle, données, calculs : quelles infrastructures dans un monde décarboné ?”

The Shift Project, 1.10.2025



Prévisions de la consommation mondiale d’électricité de l’IA de 2025 à 2035, en TWh
Source : « Artificial Intelligence and Electricity, A System Dynamics Approach », Schneider Electric, 2024

Orange Restricted

Prediction: shift in 2041

Energy
CO2
CSR (RSE): carbon-neutral activities

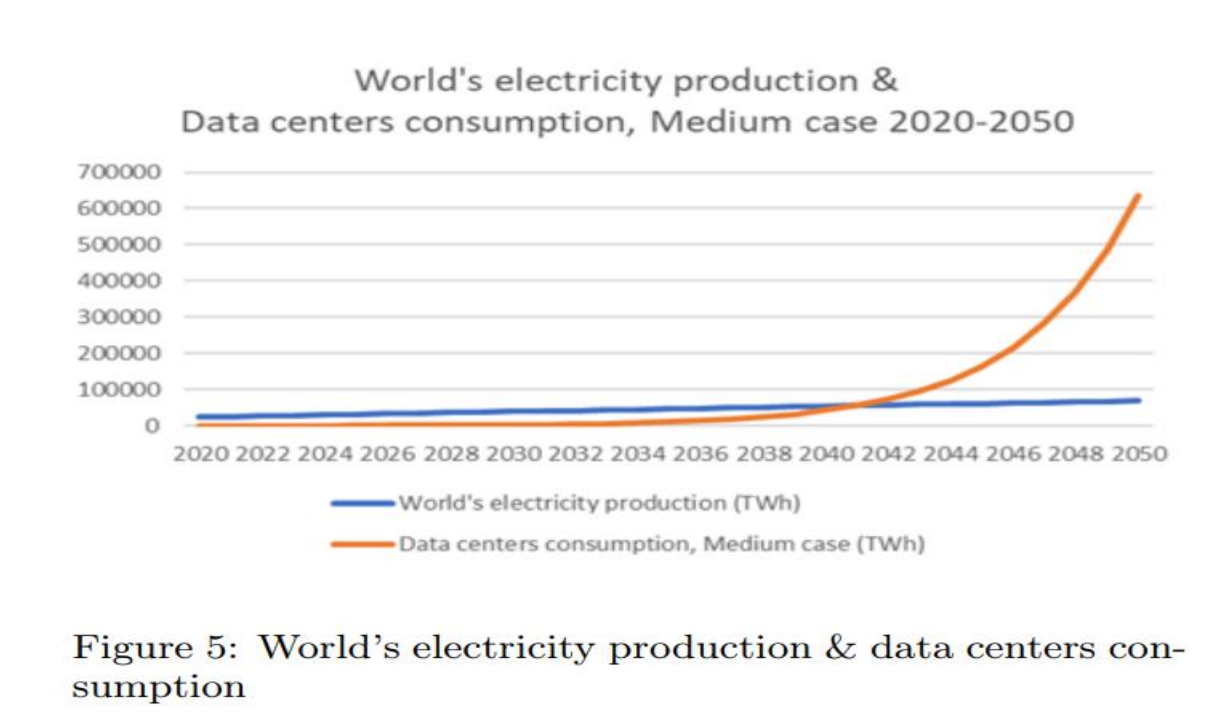
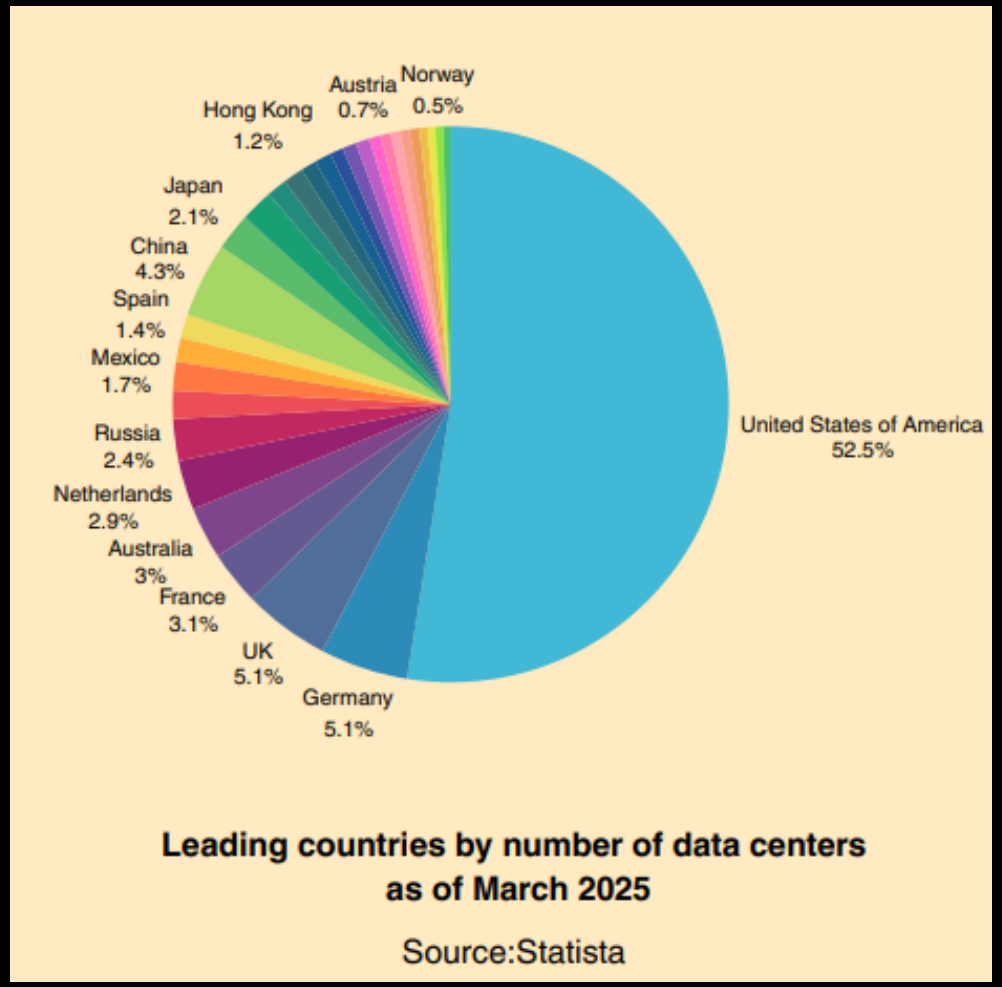


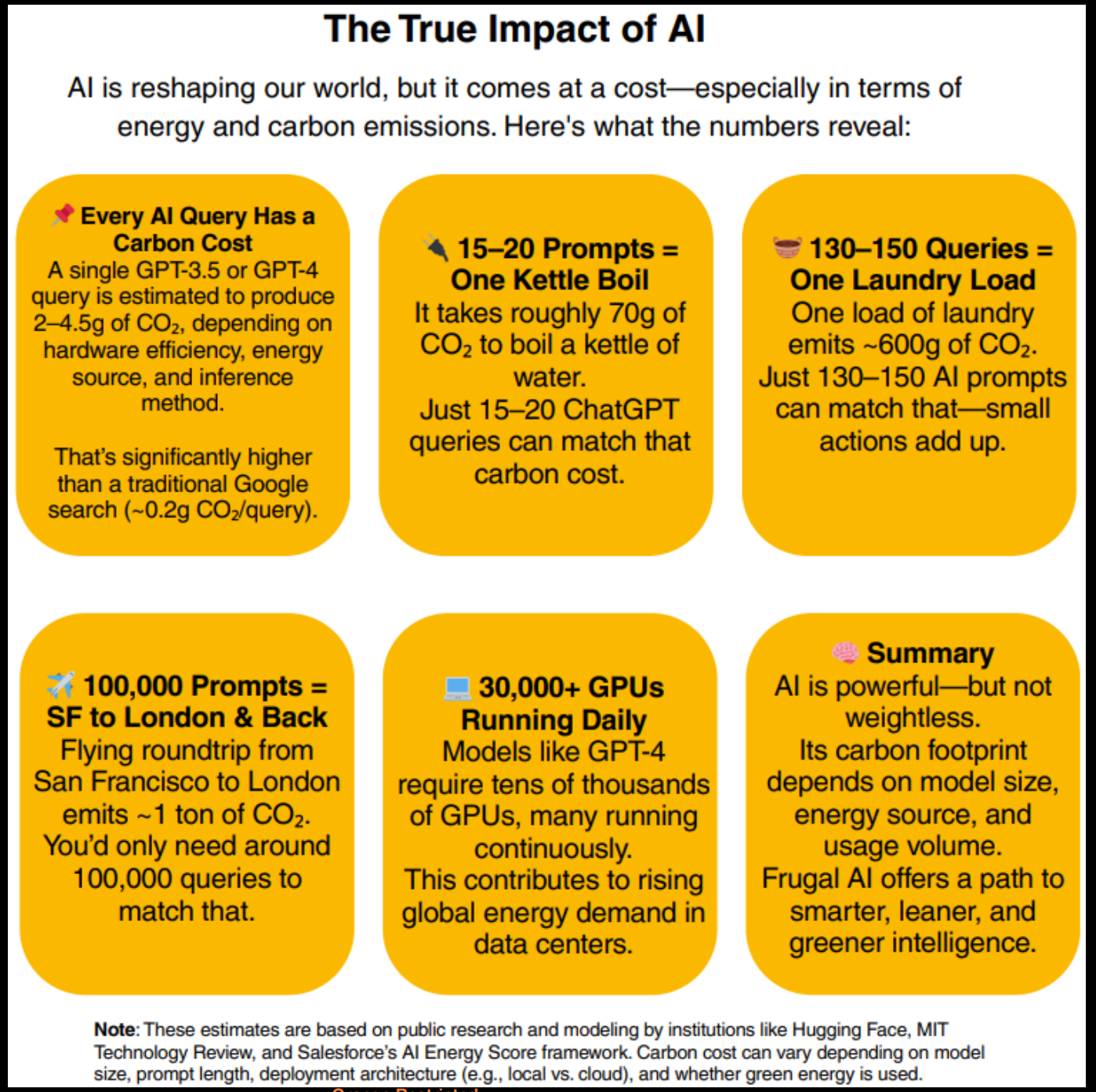
Figure 5: World's electricity production & data centers consumption



Orange Restricted

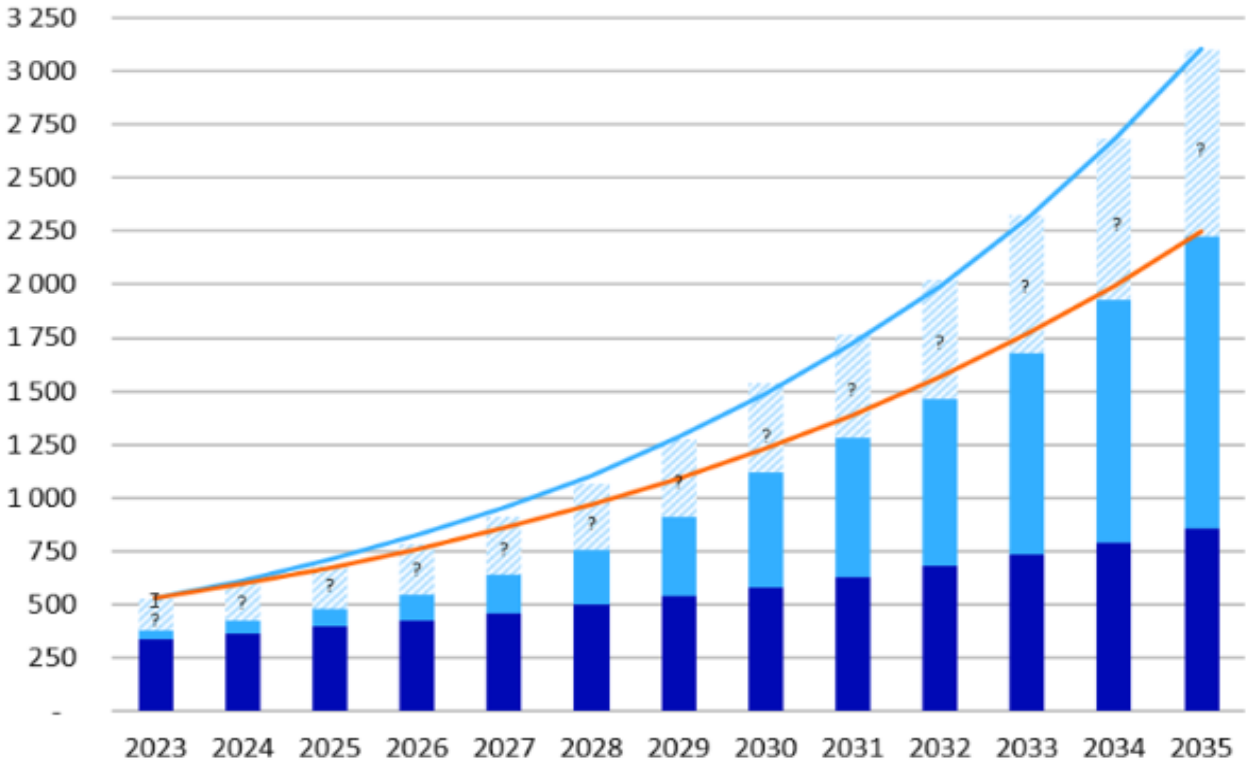
Cost of AI

"Frugal AI Hub", White Paper, Univ. of Cambridge, April 2025.



Orange Restricted

Cost of AI



Consommation mondiale d'électricité en phase d'usage (TWh).
Source : The Shift Project, 2025

Trois dynamiques :

Cryptomonnaies

« Phénomène IA générative »

Usages traditionnels : hébergement site web, réseaux sociaux, plateformes vidéo, cloud

« Intelligence artificielle, données, calculs : quelles infrastructures dans un monde décarboné ? »
The Shift Project, 1.10.2025

Orange Restricted

True innovation is not measured just by compute power but by human impact

The most powerful AI in the future will be the one that is:

efficient enough to scale,
inclusive enough to serve,
ethical enough to endure.

Frugal AI redefines progress by asking how we can design intelligence that is open, accessible

"Frugal AI Hub", White Paper, Univ. of Cambridge, April 2025

Orange Restricted

Perspectives

“Since these mysteries are beyond us, let’s pretend to be the organizer.” (J. Cocteau)

Orange Restricted

Did you say “rebound effect”?

The notion of “rebound effect”, theorized in the 1980s, has been known since 1865 as the “Jevons paradox”.

- It states that technological progress that reduces the consumption of a resource can lead to an increase in the total consumption of that resource.
- Historical example: James Watt’s steam engine increases coal consumption.

direct	Increased efficiency of a product or service leads to increased usage
indirect	Savings from cost reduction allow more revenue to be spent on other products and services
macroeconomic	Increased efficiency stimulates economic vitality, which generates general growth in consumption

Orange Restricted

Did you say “rebound effect”?

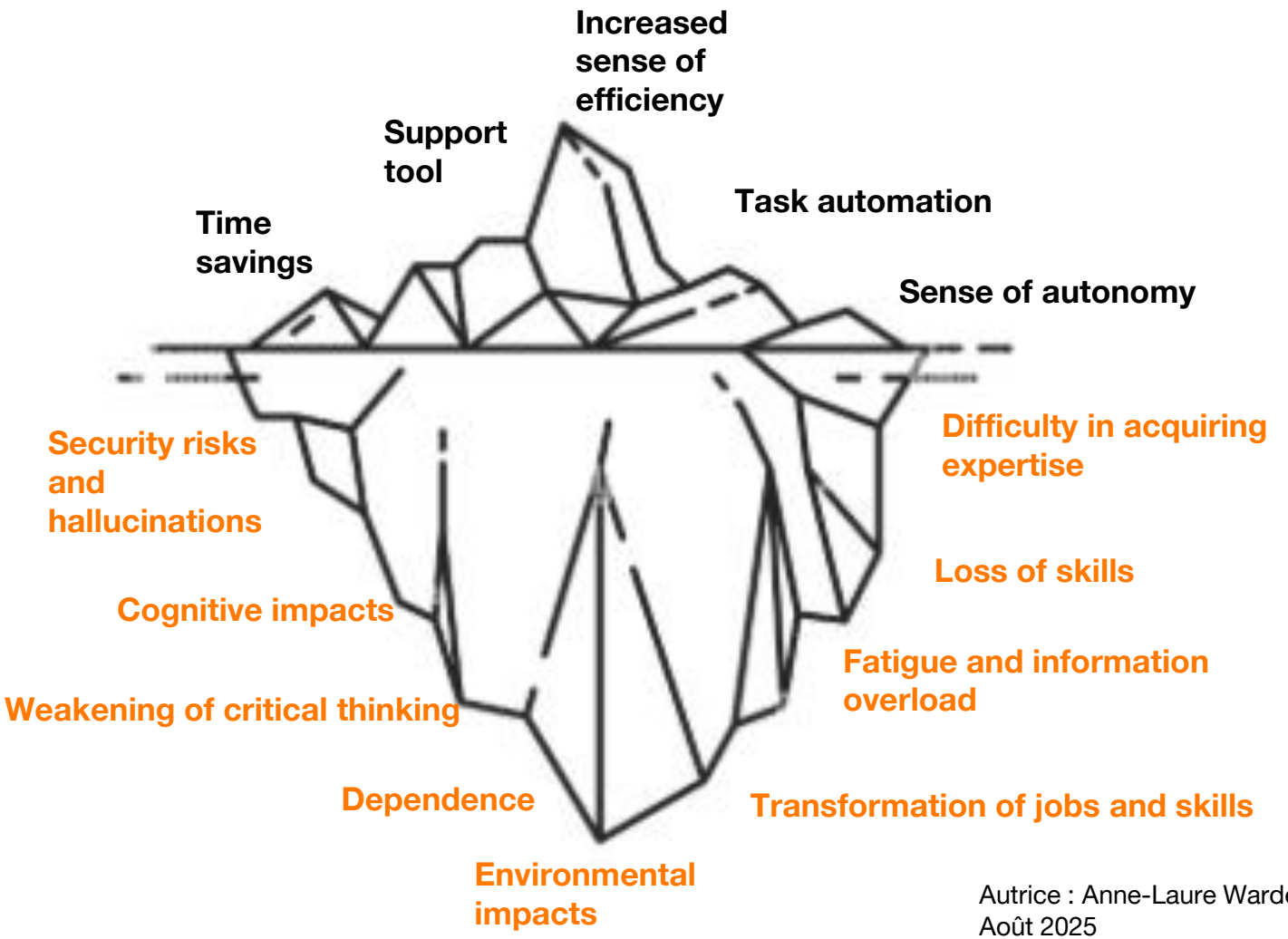
Indirect Effect	Examples for AI’s environmental impacts
Substitution impacts	Art replaced by AI-generated imagery Encyclopedias and books replaced by AI-generated content
Space rebound effects	Datacenters are getting bigger while devices are getting smaller
Scale effects	AI models growing in size and complexity Optimization approaches for AI training and deployment (batching, caching)
Direct economic rebound effects	AI hardware getting more efficient, yet datacenter energy usage is rising
Indirect economic rebound effects	More consumer devices (speakers, appliances, etc.) that allow users to interact with AI systems Smartphones replaced more often to run newest AI functionalities
Economy-wide rebound effects	Power purchase agreements from AI compute providers enable global renewable energy capacity to be expanded
Induction impacts	Targeted ads that use AI to induce more consumption
Time rebound effects	AI-optimized navigation Robots doing household chores

Table 1. Examples of AI’s indirect impacts and rebound effects, expanding on Börjesson Rivera et al [95].

From Efficiency Gains to Rebound Effects: The Problem of Jevons’ Paradox in AI’s Polarized Environmental Debate
Luccioni and al. 2025

Orange Restricted

The impacts of using generative AI in software development professions



Autrice : Anne-Laure Warde
Août 2025

Orange Restricted

AFNOR General Reference

A pioneering ADEME + AFNOR initiative to define and promote the notion of “frugal AI”:

- Around sixty companies, institutions and universities involved
- A defining effort around AI and its life cycle
- A catalog of detailed and prioritized best practices via a transparent method
- A desire to take the approach to a European scale

Orange Restricted

The models in question

The soaring parameters of LLMs:

- GPT1: 117 million parameters
- GPT2: 1.5 billion parameters (x 12)
- GPT3: 175 billion parameters (x 116)
- GPT4: 1760 billion parameters (x 10)
- GPT5: Maybe 10 x GPT4…

Size of LLMs won’t matter as much moving forward

Orange Restricted

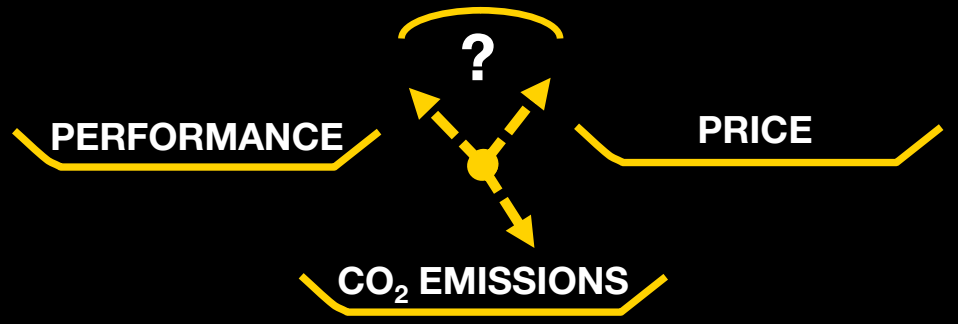
Frugal AI ?

Orange Restricted

Development criteria

A history primarily anchored:

- on performance (accuracy, response time)
- on costs



History does not stand still. A majority of actors have understood that:

- Reducing the footprint of AI is, beyond CSR, a challenge of resilience and sovereignty
- the effort is not self-evident, because economic competition induces a tyranny of speed (ttm)
- LLMs reduce the design cycle: irresistible, at the cost of certain side effects...
  - often higher operating costs, because energy-intensive models (new DeepSeek deal?)
  - explainability: the system is a black box (loss of business skills in data analysis)
  - on certain “classic” tasks: performance often slower and less precise than older models

Orange Restricted

Questioning the use of LLMs

A well-supported case study: natural language classification (sentiment analysis, etc.)

- Generative models tend to perform worse and slower
- …may be 100 times or more energy-intensive
- …and are nevertheless regularly used against older models

statistical models

Generative AI with different prompt variants

predictions

Formulas with factors of averaging and/or unit conversion are detailed in a comment attached to the cell.		Khiops	Statistical models (tested during session)			LLMs (tested outside session)			
Categories to predict		Khiops multitable	LGBMClassifier	RandomForestClassifier	LogisticRegression	gpt-35-turbo (1 few shot)	gpt-35-turbo (3 few shots)	gpt-35-turbo (9 few shots)	gpt-35-turbo (27 few shots)
Bad Flight	accuracy		na	na	na	84,66 %	88,28 %	87,41 %	85,17 %
	precision	69,00 %	83,00 %	78,00 %	71,00 %	30,31 %	32,86 %	86,44 %	82,61 %
	recall	32,00 %	78,00 %	77,00 %	74,00 %	25,86 %	44,83 %	43,97 %	32,76 %
	f1-score	43,00 %	81,00 %	77,00 %	73,00 %	40,27 %	60,47 %	58,29 %	46,91 %
Cancelled Flight	accuracy	na	na	na	na	91,90 %	92,93 %	92,76 %	91,90 %
	precision	80,00 %	82,00 %	84,00 %	81,00 %	84,16 %	82,61 %	89,36 %	90,59 %
	recall	72,00 %	84,00 %	85,00 %	80,00 %	73,28 %	81,90 %	72,41 %	66,38 %
	f1-score	76,00 %	83,00 %	85,00 %	81,00 %	78,34 %	82,25 %	80,00 %	76,62 %
Customer Service Issue	accuracy	na	na	na	na	84,83 %	86,72 %	86,38 %	86,38 %
	precision	73,00 %	82,00 %	76,00 %	79,00 %	60,14 %	65,35 %	63,70 %	64,57 %
	recall	86,00 %	74,00 %	73,00 %	72,00 %	71,55 %	71,55 %	74,14 %	70,69 %
	f1-score	79,00 %	78,00 %	75,00 %	75,00 %	65,35 %	68,31 %	68,53 %	67,49 %
Late Flight	accuracy	na	na	na	na	76,38 %	82,59 %	80,00 %	76,72 %
	precision	73,00 %	76,00 %	78,00 %	76,00 %	45,29 %	54,01 %	50,00 %	45,85 %
	recall	71,00 %	84,00 %	78,00 %	75,00 %	87,07 %	87,07 %	87,93 %	90,52 %
	f1-score	72,00 %	80,00 %	78,00 %	75,00 %	59,59 %	66,67 %	63,75 %	60,87 %
Lost Luggage	accuracy	na	na	na	na	92,59 %	93,45 %	92,93 %	93,62 %
	precision	79,00 %	86,00 %	88,00 %	84,00 %	95,06 %	94,32 %	95,18 %	95,40 %
	recall	75,00 %	88,00 %	91,00 %	90,00 %	66,38 %	71,55 %	68,10 %	71,55 %
	f1-score	77,00 %	87,00 %	90,00 %	87,00 %	78,17 %	81,37 %	79,40 %	81,77 %
Average	accuracy	**82,00 %**	**82,00 %**	**81,00 %**	**78,00 %**	**86,07 %**	**88,79 %**	**87,90 %**	**86,76 %**
	precision	**74,80 %**	**81,80 %**	**80,80 %**	**78,20 %**	**75,11 %**	**77,83 %**	**76,94 %**	**75,80 %**
	recall	**67,20 %**	**81,60 %**	**80,80 %**	**78,20 %**	**64,83 %**	**71,38 %**	**69,31 %**	**66,38 %**
	f1-score	**69,40 %**	**81,80 %**	**81,00 %**	**78,20 %**	**64,34 %**	**71,81 %**	**69,99 %**	**66,73 %**
energy cost for 1 vectorization (Wh) (obtained by averaging 29000 repetitions)		na	3,168E-04	3,168E-04	3,168E-04	na	na	na	na
energy cost for 1 model prediction (Wh) (obtained by averaging ~10^6 transforms)		3,329E-06	9,227E-08	1,845E-07	1,475E-08	na	na	na	na
total energy cost for 1 inference (Wh) for statistical models: vectorization + transform for LLMs: prompting estimation via ecologits		3,329E-06	3,169E-04	3,170E-04	3,168E-04	2,690E-02	2,690E-02	2,690E-02	2,690E-02

energy behavior in inference

Model	F1	Energy (Wh)
Khiops	69.4	3.3 E-06
Random Forest	81	3.2 E-04
GPT 4-o mini	74.7	1.6 E-02

Orange Restricted

Transition “under constraint” = pragmatism

- Being aware of your limits and vulnerabilities
- Questioning the need: (re)learning to distinguish the necessary from the superfluous
- Choosing the right model for the right use
- Resistance to the “environment”: “you won’t be able to do without it…” (inductivist turkey!)

1876

we won’t be able to do without it!

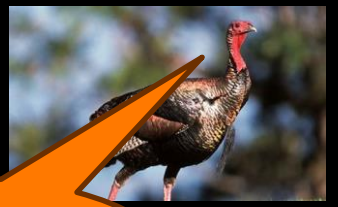
1895

we won’t be able to do without it!

1980

we won’t be able to do without it!

1991



we won’t be able to do without it!

2022

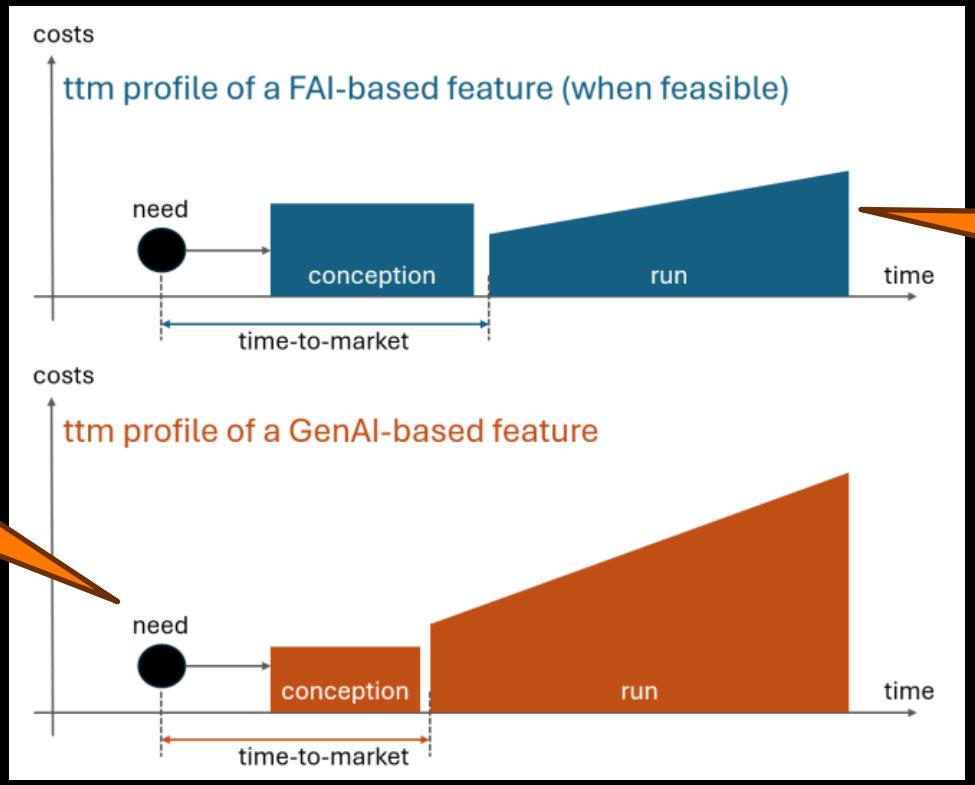
?

Orange Restricted

Transition: cooperation vs competition?

- Resistance towards …oneself : rebound effect + pressure of “time-to- market”

fast dev, energy-intensive, expensive simple tasks: mediocre



slow dev, “ad hoc” sober, economical, explainable

https://hal.science/hal-05049765v1/file/Frugal_AI_Introduction_Concepts_Development_and_Open_Questions_HAL.pdf

- We have (for now) the luxury of being able to ask ourselves the question...

Orange Restricted

Frugal AI
What is it?

- Frugal AI integrates artificial intelligence with principles of frugality, focusing on minimal resource use while minimising functionality, robustness, and accessibility.

- It contrasts with efficiency by emphasizing simplicity, affordability, and sustainability from the outset, rather than optimizing existing systems.”

Orange Restricted

Domains of application

- Customer Experience: efficient chatbots and virtual assistants (e.g., using FrugalGPT approaches)
- Industry: Building robust machine learning models for predictive maintenance or process optimization on the shop floor, even with limited or variable data quality, running on edge devices.
- Healthcare: lightweight diagnostic tools deployable on mobile devices for disease detection (e.g., image analysis) in remote or under-resourced areas
- Education: Creating cost-effective AI tutoring systems accessible on low-end devices
- Environment: Employing AI for optimizing energy distribution in microgrids; analyzing environmental data (e.g., pollution, deforestation) from limited sensor networks
- Agriculture: Enabling precision agriculture solutions (e.g., crop monitoring, disease detection) using AI on low-power devices or analyzing limited drone/satellite imagery in remote farming communities.

Orange Restricted

Social definition vs regulatory definition(s) of AI systems

- Users
  - Will Heaven in the MIT Technology Review: “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.”
  - Hubert Guillaud: AI is the set of techniques that stand between lab research and widespread usages.
  - AI was used to describe image recognition
  - In 2025, AI is identified as LLMs (Large Language Models)

Important to educate users/populations to keep critical thinking in AI usages
-> avoid replacing prompts for questions and feedback for answers

- AI system [EU AI Act, Article 31] 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 229892]: AI system that accumulates, combines and encapsulates knowledge provided by a human expert or experts in a specific domain to infer solutions to problems.

Orange Restricted

Strategy for Frugality ?

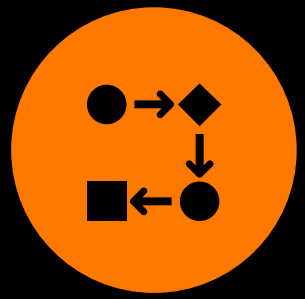
Orange Restricted

Frugal AI - Key principles (First sight)

Efficiency
Prioritization of 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.

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.

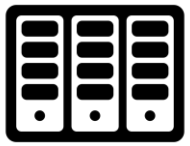
Simplicity
Frugal AI solutions often prioritize simplicity and ease of use over complexity.
This may involve using simpler algorithms or user interfaces that require less training and technical expertise to operate. Efficient workflows also support the frugal mindset.

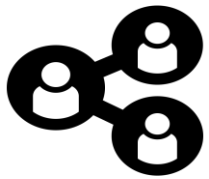


Scalability
Frugal AI solutions should be scalable, allowing them to adapt to different contexts and user needs without significantly increasing costs.
This may involve designing modular architectures that can be easily expanded or customized as needed.

Orange Restricted

Frugal AI - Key principles (Second sight)

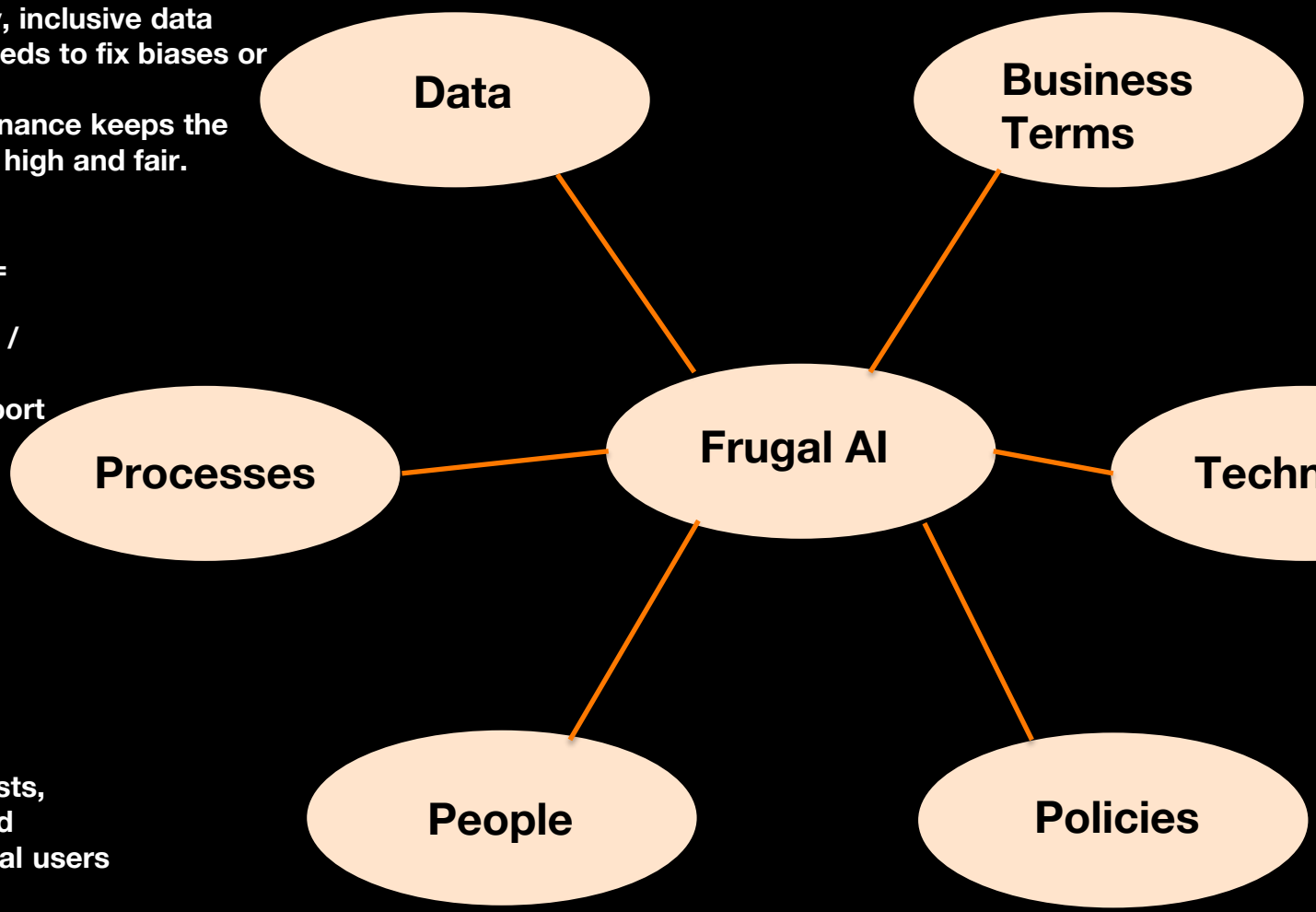




- Resource Efficiency throughout the AI lifecycle: computational power, energy, data, financial capital
  - employing optimized or low-cost algorithms
  - optimizing models for size and speed, reusage
  - practicing data minimization
- Sustainability: ecological goals and net-zero targets
  - reducing the carbon footprint of AI through energy-efficient model training and inference
  - promoting the use of sustainable data centers powered by renewable energy
  - utilizing tools like AI Energy Scoring to benchmark and incentivize lower energy consumption
- Accessibility & Inclusion: democratize access to AI capabilities, lower technical and financial barriers
  - diverse user groups
  - designing lightweight models deployable on low-power edge devices (smartphones, IoT sensors) and in resource-constrained environments
- Impact & Scalability: reach a wider audience quickly
  - social, environmental, or economic
  - optimizing models for low-cost deployment and leveraging lightweight architectures
  - rapid scaling of solutions without massive infrastructure investments

Orange Restricted

Designing, developing, deploying AI systems that are resource-efficient, affordable, accessible and sustainable



High-quality, inclusive data excludes needs to fix biases or errors later.
Good governance keeps the data quality high and fair.

Clear definitions help align AI efforts with real business goals: delivering cost-effective, targeted solutions that make a genuine impact.

Streamlined processes = less waste and more efficient AI development / deployment.
Efficient workflows support the frugal mindset.

Scalable, resource-efficient technology is the key. Governance helps pick the right tech to balance cost-effectiveness and quality.

Data scientists, analysts, and business/final users

Provide the ethical and operational guardrails needed for data and AI use, aligning efforts with resource efficiency and compliance

Orange Restricted

Strategies for Frugality
Key findings

Eco-Design Principles: Prioritize traditional AI over generative AI, measure environmental impacts throughout the lifecycle, and optimize functionalities to essential needs.

Resource-Aware Design: Adapting AI systems to available resources.

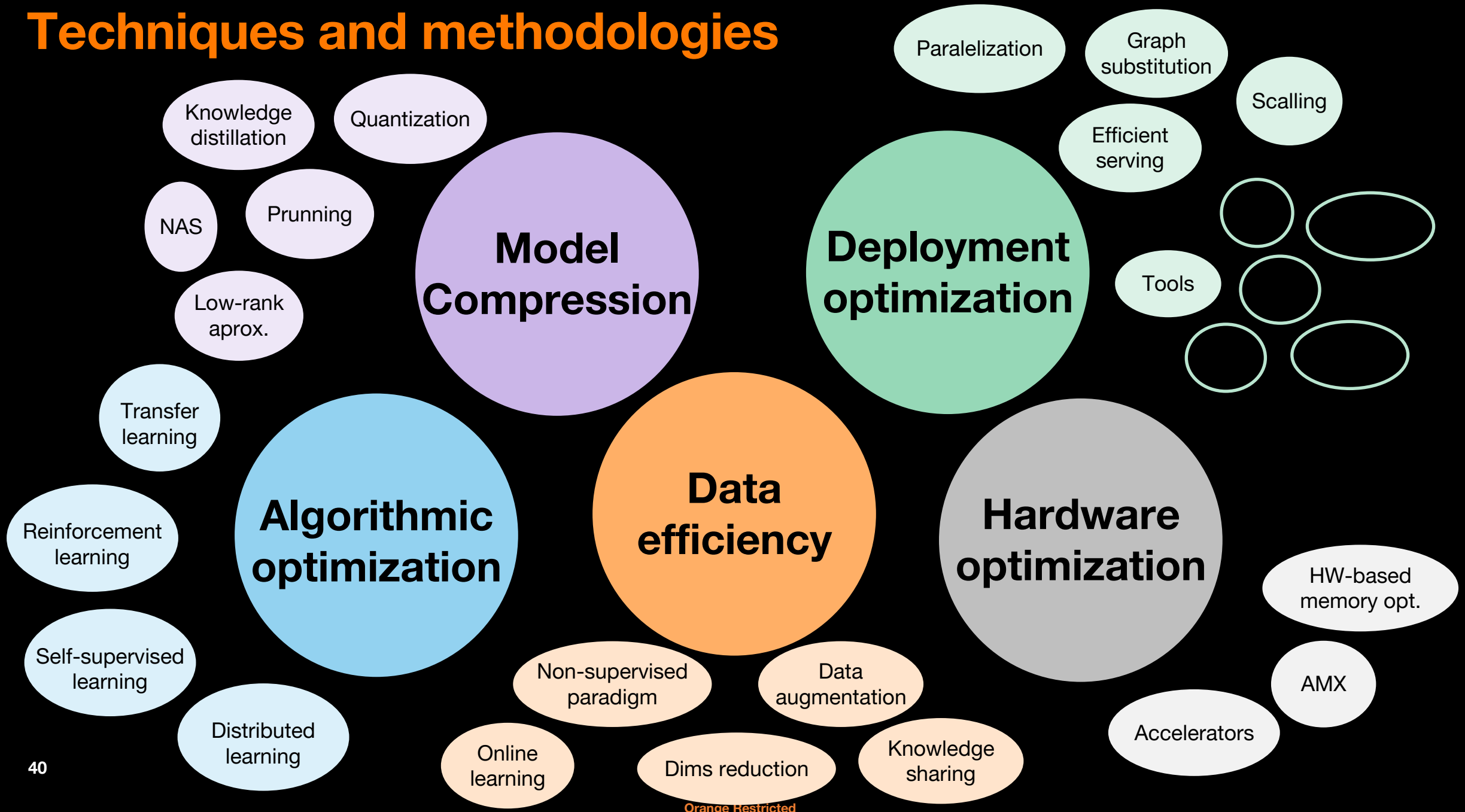
Reuse and Fine-Tuning: Utilizing pre-trained models to reduce training costs.

Model Compression: Reducing model size and complexity while maintaining the necessary performance.

Hardware Optimization: Leveraging specialized hardware for energy efficiency. Advances in low-power GPUs, neural processing

Orange Restricted

Techniques and methodologies



Orange Restricted

Hardware ecosystem

Efficient Architectures & Hardware Optimization

Uses AI models designed for efficiency (e.g., MobileNet, EfficientNet).

Optimized execution using specialized AI chips (TPUs, FPGAs).

Example: EfficientNet achieves state-of-the-art performance with 5x fewer computations.

Edge AI & TinyML

Runs models directly on low-power hardware like sensors or microcontrollers.

Minimizes latency, energy use, and data transfer costs.

Example: Wake word detection on a $2 microcontroller with <50KB RAM.

Low-Power Hardware

Uses AI chips optimized for energy efficiency (e.g., TPUs, neuromorphic processors).

Delivers faster, greener inference at the same or lower cost.

Example: Google's Edge TPU enables 4 TOPS with <2W power draw.

Neurosymbolic & Swarm Intelligence

Combines logic-based reasoning with learning or distributed agent coordination.

Enables lightweight, explainable, and distributed intelligence.

Example: Neurosymbolic AI in robotics reduces data dependency while improving interpretability.

Sustainable Data Centers

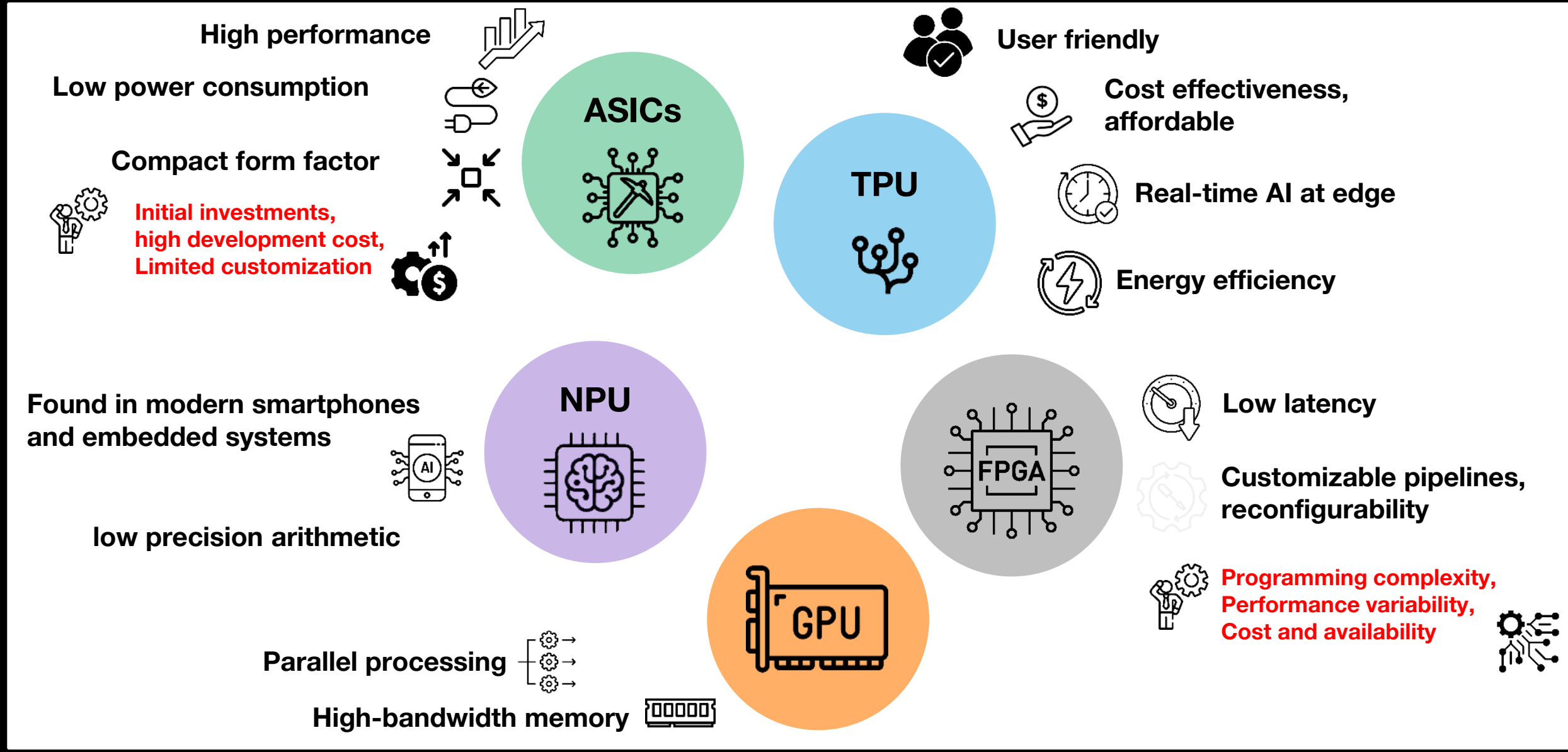
Deploys AI workloads in facilities powered by renewables & efficient cooling.

Supports green innovation goals and long-term TCO reduction

New advancements are also brought by using AI to design energy efficient data centres.

Orange Restricted

Hardware



Orange Restricted

Techniques and methodologies

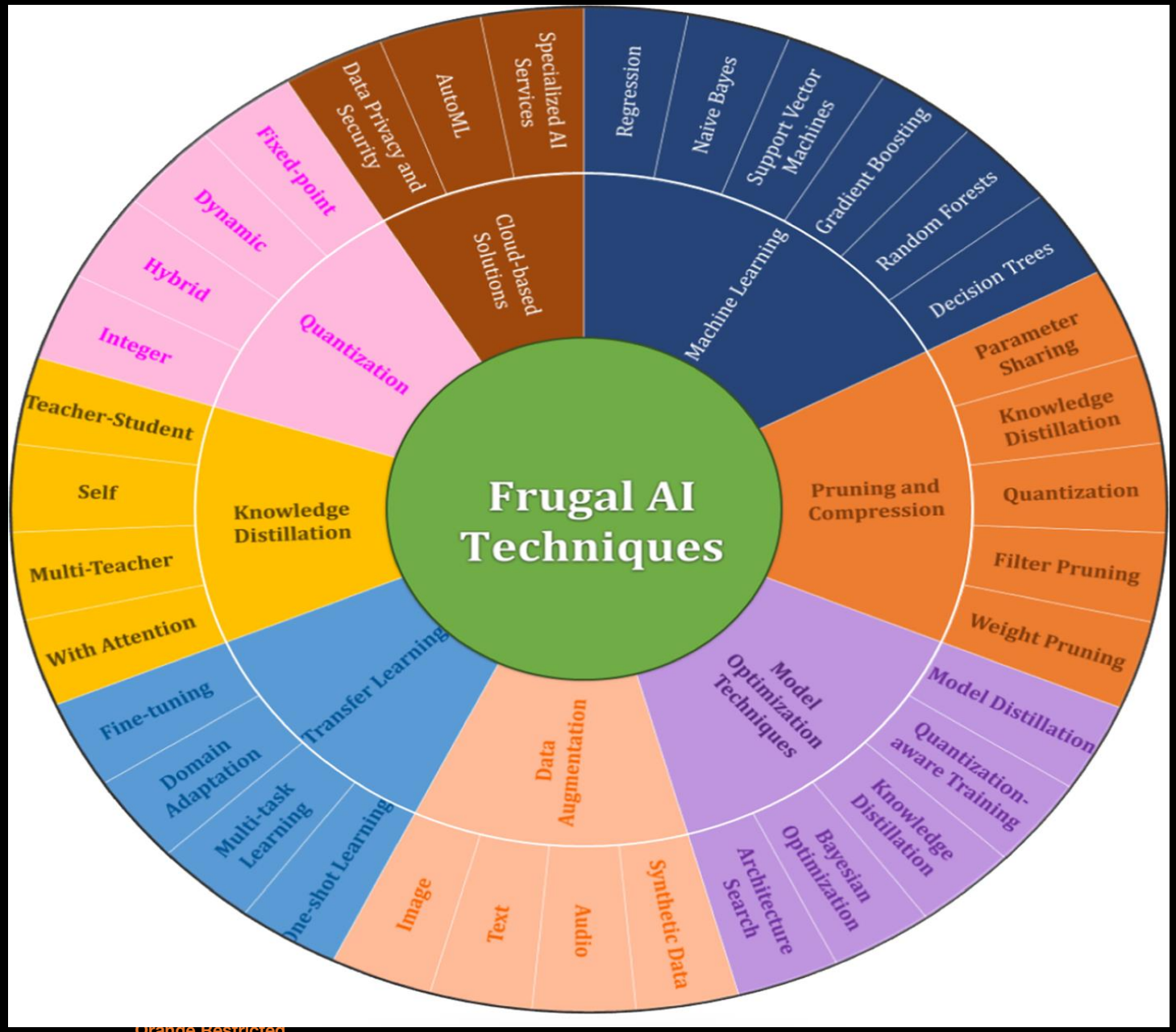


Figure from
N. Girdhar *et al.,* A comprehensive review of frugal AI: challenges, applications, and the road to sustainable AI. *Soft Comput* 29, 4823–4856 (2025).

Orange Restricted

Measuring Frugality ?

Orange Restricted

Measuring frugality (1/2)

Frugal AI Self-Assessment Framework: evaluate maturity across key dimensions of Frugal AI.
Structured way to identify strengths and areas for improvement: sustainable data centers, AI efficiency techniques, energy scoring, handling constraints, ROI/TCO calculation, scalability, ecosystem engagement, inclusivity, and impact focus.

AI Energy Scoring: measuring and benchmarking the energy consumed by AI models during training and inference relative to their performance (e.g., accuracy, latency).
Guiding developers towards more sustainable choices: compare models based on their energy efficiency by tools impact calculators (Code Carbon, ML CO2, etc.) and standardized scoring systems

Standardization (e.g., AFNOR Spec 2314): concrete benchmarks and guidelines for evaluating the environmental impact of AI services.

Encourage transparency and accountability by requiring justification for

- using AI over less intensive alternatives,
- adherence to best practices throughout the lifecycle, and
- consideration of planetary boundaries.

Further development and international adoption of such standards

Orange Restricted

Measuring frugality (2/2)

Value Clarification Matrix: prioritize AI projects by assessing them against multiple criteria

A holistic view beyond pure technical performance: Business Value (revenue/cost impact), Environmental Impact (resource reduction), Social Impact (accessibility/inclusion), Implementation Risk, and Total Cost of Ownership (lifetime resource needs).

Total Cost of Ownership (TCO) & Return on Investment (ROI): moving beyond basic ROI calculations for demonstrating the economic benefits of frugal approaches

Over the AI system's lifecycle, include

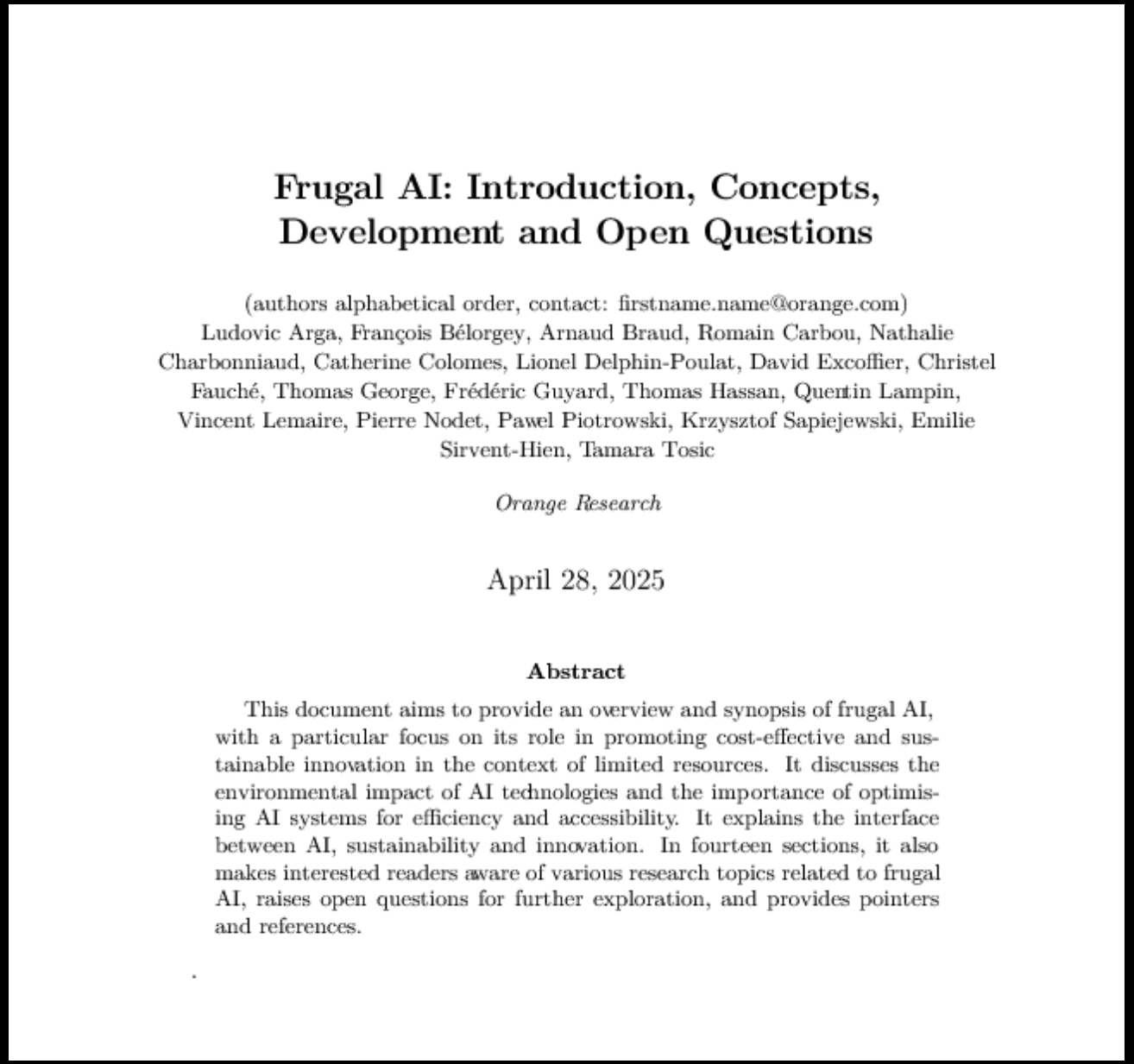
- initial development costs
- ongoing infrastructure, compute, energy, data storage, and maintenance expenses.

Orange Restricted

One of our contribution

Orange Restricted

Frugal AI
Research paper



Frugal AI: Introduction, Concepts, Development and Open Questions

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

Orange Research

April 28, 2025

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.

orange™

Orange Restricted

Some words

What is the purpose of the document?

- The aim of this document is not to go into frugal AI in depth, but simply to provide an introduction, give a few related topics, give some pointers, provide references, raise awareness. Frugal Innovation is an opportunity to innovate in a cost effective and sustainabe way under resource scarcity.
- The document is considered a valuable resource for readers interested in understanding the intersection of AI, sustainability, and innovation.

Orange Restricted

What and for whom?

General public / society

- Raise public and organizational awareness about the environmental impact of AI and frugal practices.

Researchers

- Support interdisciplinary research on frugal AI and its applications
- Contribute to clear standards and benchmarks for measuring the environmental footprint of AI systems
- Open questions

Managers / stakeholders

- Need to build new bussines models that take into account all the frugal measures
- Resource consumption measurements
- Right AI for the right need at the right time

Developers / Data Scientists

- Optimization techniques
- Taxonomy of tools, benchmarks, frameworks and many sources for more in-depth knowledge of each topic

Orange Restricted

Some words
findings

- 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.

Orange Restricted

About this document

Frugal AI: introduction, concepts, development and open questions

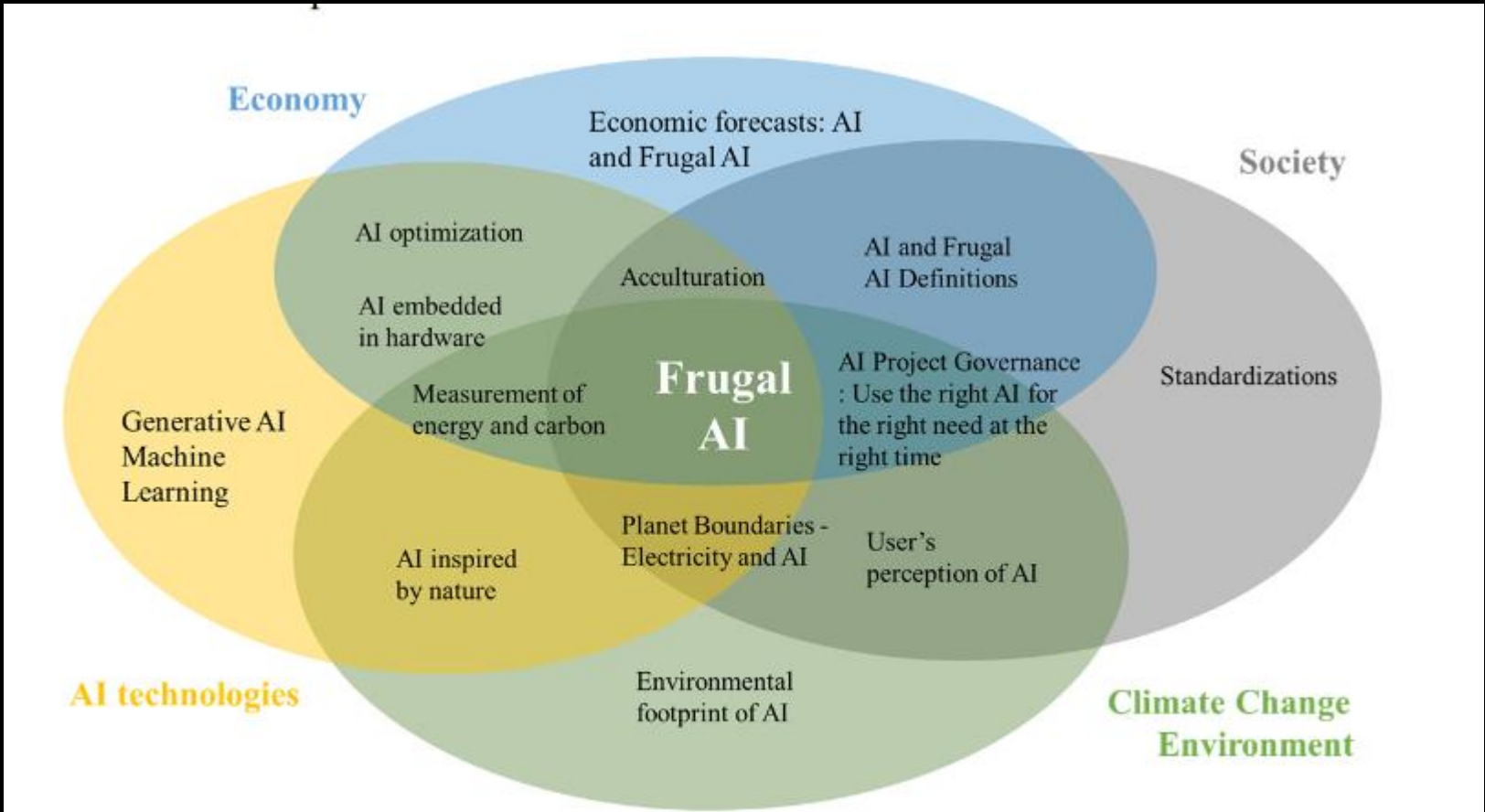


Figure 1: Frugal AI: what impact and what solutions for the environment?

Orange Restricted

About this document

Frugal AI: introduction, concepts, development and open questions

Less than 20% of companies achieve sustainable Return on Investment (ROI) from AI: high costs and immature Total Cost of Ownership (TCO) methodologies

Large models consuming energy equivalent to hundreds of homes and contributing significantly to carbon emissions

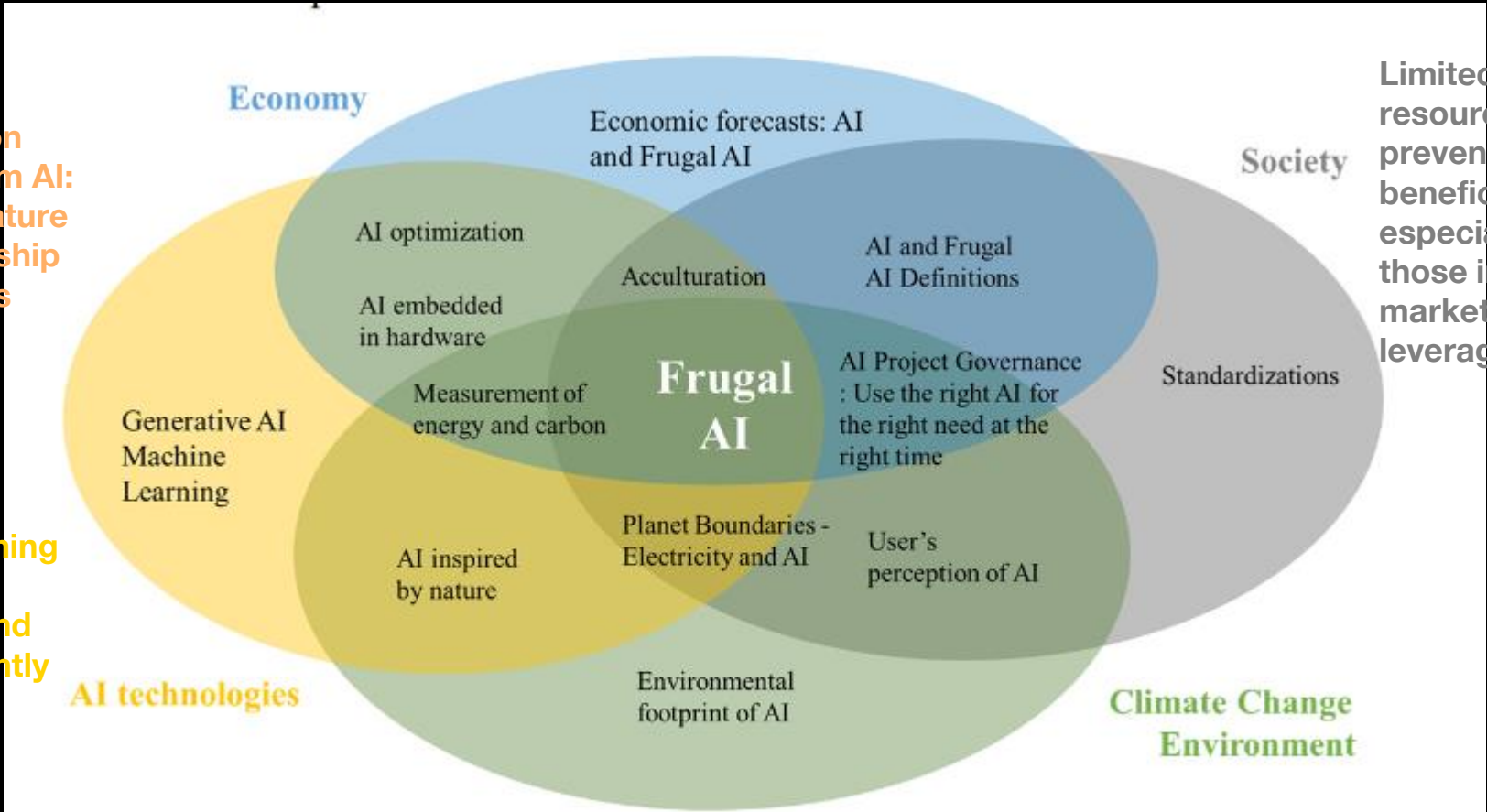


Figure 1: Frugal AI: what impact and what solutions for the environment?

Limited Reach: resource constraints prevent many potential beneficiaries, especially SMEs and those in emerging markets, from leveraging AI

Orange Restricted

Environmental Footprint of AI
Key findings

AI contributes significantly to greenhouse gas emissions, with digital technologies accounting for nearly 4% of global emissions. Generative AI exacerbates this impact due to high energy and water consumption, as well as resource-intensive data center operations. AI's energy and resource demands (electricity, water, rare metals) are expected to grow exponentially, raising sustainability concerns. However, AI also offers potential for environmental monitoring and efficiency improvements in various sectors.

Opportunity and Challenge

- **Generative AI's growing demand for computational power and data center capacity poses challenges for electricity and resource availability.**
- **Solutions like renewable energy integration and hardware innovations are critical.**
- **Explore Frugal AI application in critical sectors (eg : telecom, agriculture, healthcare...)**

Orange Restricted

Economic and societal implications
Key findings

Economic Trends: AI adoption is growing rapidly, driven by both professional and consumer demand. However, high operational costs of large language models (LLMs) may push organizations toward more frugal alternatives.

Public Perception: Awareness of AI is high, but its environmental impact is largely invisible in public debates.

Regulatory: There is a strong demand for regulatory frameworks to ensure ethical and sustainable AI deployment.

Opportunity and Challenge

- **Frugal AI can drive innovation in underserved regions, fostering accessibility and inclusion (eg: on the African market)**
- **However, societal debates on AI's ethical, environmental, and economic impacts remain polarized.**
- **Contribute to the standards for measuring AI's environmental footprint.**

Orange Restricted

Strategies for Frugality
Key findings

Eco-Design Principles: Prioritize traditional AI over generative AI, measure environmental impacts throughout the lifecycle, and optimize functionalities to essential needs.

Resource-Aware Design: Adapting AI systems to available resources.

Reuse and Fine-Tuning: Utilizing pre-trained models to reduce training costs.

Model Compression: Reducing model size and complexity while maintaining the necessary performance.

Hardware Optimization: Leveraging specialized hardware for energy efficiency. Advances in low-power GPUs, neural processing units (NPUs), and application-specific integrated circuits (ASICs) enable energy-efficient AI applications, particularly in edge computing.

Opportunity and Challenge

- **Develop models compression and hardware optimizations to improve AI efficiency.**
- **Explore frugal AI applications in telecom automation, and semantic communication.**
- **Investigate the potential of nature-inspired AI and emerging technologies (e.g., quantum and neuromorphic computing) for sustainable innovation.**

Orange Restricted

Where are we? Challenges and open questions (1/2)

Feature	Frugal AI	Classical AI
Resources	Minimal compute, energy, data, light models	High compute, energy, data requirements, complex models
Sustainability	lower carbon footprint by design	More carbon-intensive
Cost		High infrastructure and operational costs
Accessibility	Democratized	Often elite-oriented
Performance	Efficiency-optimized	State-of-the art
Data	Quality and relevance for effectiveness with less data	Massive dataset volumes
Deployment speed	Faster prototyping/deployment	May include long cycles
Design	Constraint driven	Performance maximisation

Unified tools
CO2 transparency
Measures

Automatized designs
Choice of frugal algo
Memory

RSE requirements
RSE reglementation

Strategy

Data/Model Democracy

Users / Customers

Orange Restricted

Where are we? Challenges and open questions (2/2)

Does reusability/fine-tuning/sparse learning/ make AI frugal? Will advancing learning theory result in more frugal AI models?

Should AI be resource-aware to be frugal? Frugal AI as a desirable side-effect of resource-constrained innovation? Frugal continuous learning?

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

What social usages could

Can complex scalable systems be conceived as Frugal by design Will very large generative AIs (LLMs) and their uses one day become frugal? Is semantic communication a means to frugal Agentic communications?

What about the future of AI in a constrained environment?

What could be frugal in telecom network automation?

Orange Restricted

Steps towards Frugal AI

Orange Restricted

Steps towards Frugal AI

- Understand the impact that our AI system has on our company, society and planet
- Apply eco-design on our AI use-case
- Understand the alternative setups with limited resources
- Conceive our AI for current needs and usages (final training model and its intermediate steps),
- Apply recommendations, specifications and regulations

Orange Restricted

Discussions / open research questions (1/2)

How can telecommunication networks efficiently support semantic communications between autonomous agents at scale?

Can we develop specific encodings for semantic representations, similar to audio and video codecs optimization of 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?

Orange Restricted

Open research questions (2/2)

Does reusability/fine-tuning/sparse learning/… make AI frugal? Will advancing learning theory result in more frugal AI models?

Should AI be resource-aware to be frugal? Frugal AI as a desirable side-effect of resource-constrained innovation? Frugal continuous learning?

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

What social usages could bring to the frugal AI questioning?

Can complex scalable systems be conceived as Frugal by design? Will very large generative AIs (LLMs) and their uses one day become frugal? Is semantic communication a means to frugal Agentic communications?

What about the future of AI in a constrained environment?

What could be frugal in telecom network automation?

Orange Restricted

There is no single frugal approach that fits all the needs ?

Orange Restricted

What do we do at Orange?
(additionally, to the research paper)

Orange Restricted

4 pillars

Governance

Measurement

Acculturation

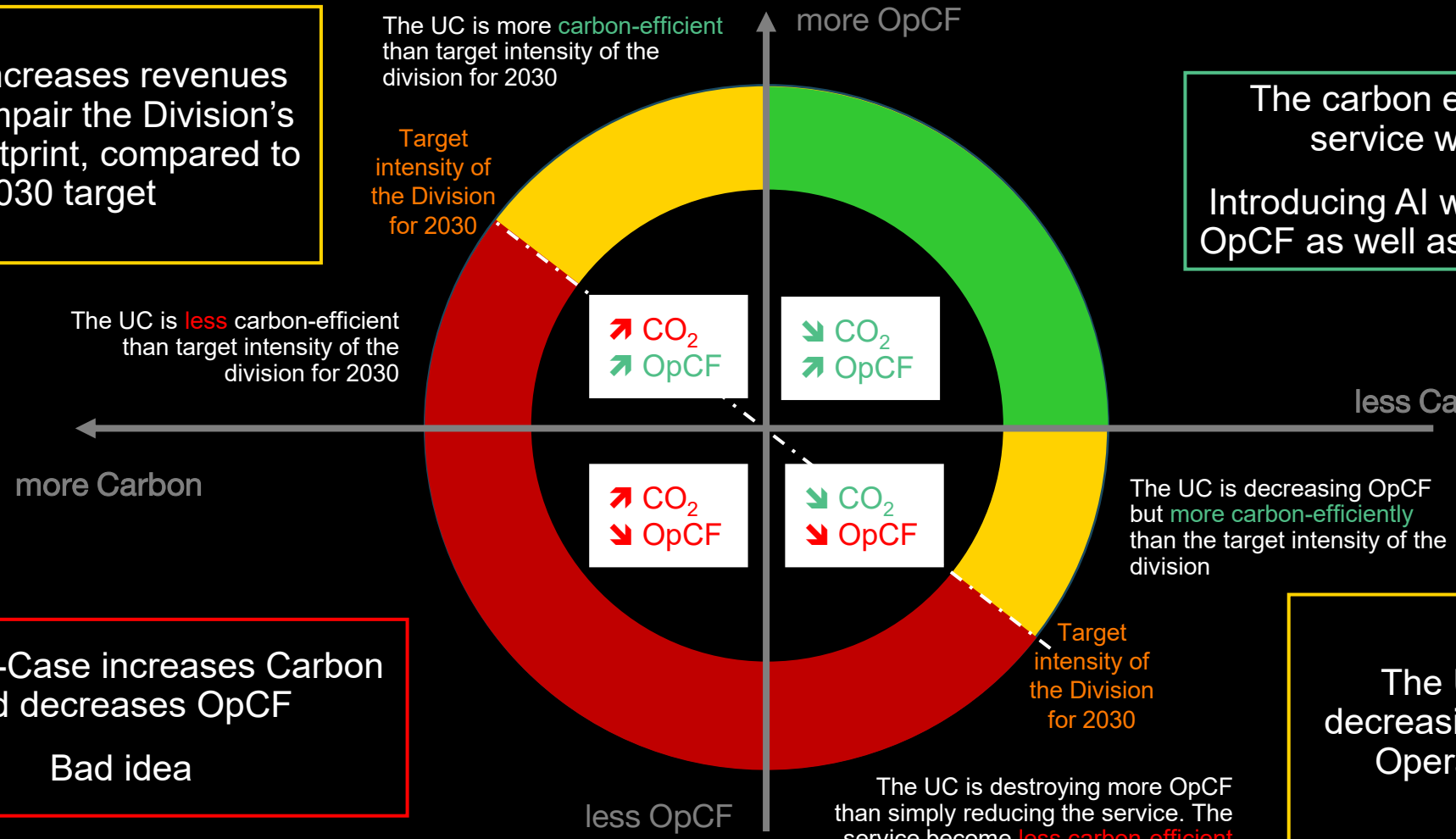
Research

Orange Restricted

Governance

AI Carbon index - How it helps for decision

The UC increases revenues but may impair the Division's carbon footprint, compared to 2030 target



The carbon efficiency of the service will improve.

Introducing AI will generate more OpCF as well as decrease carbon

The Use-Case increases Carbon and decreases OpCF

Bad idea

The Use-Case is net decreasing Carbon but also Operating Cash-Flow

Orange Restricted

Deuxième partie

Orange Restricted

Définir le juste nécessaire technologique

Principe : choisir la solution la plus simple et la plus sobre qui réponde efficacement au besoin.
Une solution est légitime si :

- **elle répond à un besoin réel,**
- **elle est proportionnée à l'usage,**
- **elle apporte une valeur claire,**
- **son coût et son impact sont maîtrisés.**

Elle est utile mais dispensable si :

- **elle améliore l'expérience,**
- **sans répondre à une nécessité forte,**
- **et qu'une solution plus simple reste possible.**

Elle est à éviter si :

- **elle est surdimensionnée,**
- **son impact est disproportionné,**
- **ou elle est moins pertinente qu'une alternative plus simple.**

Critères de décision -Pertinence

- **Performance utile**
- **Coût total**
- **CO2/énergie**
- **Simplicité**
- **Robustesse**
- **Effets sociaux**

Le juste nécessaire technologique, c'est la technologie suffisante pour répondre utilement à un besoin réel, sans excès de complexité, de coût ni d'impact.

Orange Restricted

Pour mon périmètre l'AI
De quoi parle-t-on ? Quel périmètre ?

Le numérique ne se limite pas à l'outil visible par l'usager.

Il faut considérer tout le cycle de vie du service numérique ou d'un système d'IA :

- **conception**
- **collecte et stockage des données**
- **entraînement / paramétrage**
- **inférence / usage**
- **infrastructure (terminaux, réseaux, data centers)**
- **maintenance / mises à jour**
- **fin de vie**

J'insiste sur une approche cycle de vie et non sur le seul usage visible.

Orange Restricted

Quels critères d’impact regarder ?

Critères techniques et économiques

- performance utile
- coût de développement
- coût d’exploitation
- coût total sur le cycle de vie (TCO)
- rapidité de déploiement
- robustesse / maintenabilité

Critères environnementaux

- consommation d’énergie
- émissions de CO2
- consommation d’eau
- usage de ressources matérielles / métaux / serveurs
- déchets électroniques
- impacts du data center et de l’infrastructure

Critères sociaux et organisationnels

- accessibilité / inclusion
- effets sur les compétences
- perte d’expertise
- dépendance
- fatigue informationnelle
- affaiblissement de l’esprit critique
- transformation du travail
- acceptabilité sociale

Orange Restricted

Critères souvent oubliés à relancer

Ne pas s’arrêter aux seuls critères classiques (prix / performance).

On peut relancer sur les critères moins spontanément cités :

- eau consommée
- fabrication du matériel
- fin de vie des équipements
- effet rebond (un gain d’efficacité peut conduire à plus d’usages et donc plus d’impact)
- coûts cachés de maintenance
- coûts humains (formation, supervision, contrôle, correction)
- dégradation des compétences
- niveau réel de besoin
- existence d’une alternative plus simple

👉 ce qui n’est pas visible n’est pas forcément négligeable.

Orange Restricted

Qu’est-ce qui est “juste” dans mon contexte ?

Une solution « me » semble “juste” si elle est :

- adaptée au besoin réel
- proportionnée à l’enjeu
- la plus simple possible
- suffisamment performante, sans surdimensionnement
- soutenable économiquement
- sobre en ressources
- acceptable socialement
- gouvernable dans la durée

On peut se poser la question :

- Quel est le besoin réel ?
- Faut-il vraiment une solution numérique ?
- Si oui, faut-il vraiment de l’IA ?
- Si oui, faut-il vraiment de l’IA générative ?
- Quelle solution plus simple pourrait suffire ?
- Quel niveau de performance est suffisant, pas maximal ?
- Quels impacts sommes-nous prêts à accepter ?
- Quels impacts nous paraissent disproportionnés ?

Orange Restricted

Quels besoins rendent le numérique nécessaire ?

Le numérique devient nécessaire lorsqu'il permet de répondre à un besoin que des moyens non numériques, ou moins numériques, ne peuvent pas satisfaire de manière suffisante, soit pour des raisons :

- d'échelle,
- de vitesse,
- de complexité,
- de coordination,
- de traitement de données,
- ou d'accessibilité.

Autrement dit, le numérique est légitime non parce qu'il est moderne, mais parce qu'il est irremplaçable ou nettement supérieur pour remplir une fonction donnée.

Orange Restricted

Frugalité vs Sobriété ?

Point commun : réduire les impacts du numérique

Objectif commun :
- limiter le gaspillage,
- réduire la consommation de ressources,
- contenir les impacts environnementaux,
- éviter les usages numériques disproportionnés.

Elles s'inscrivent donc toutes les deux dans une critique d'un numérique fondé sur :
- la surconsommation,
- la course à la performance,
- l'inflation des modèles,
- et le déploiement d'outils sans interrogation suffisante du besoin réel.

Orange Restricted

la sobriété agit surtout sur la consommation
la frugalité agit dès la conception

Sobriété = consommer moins ?
Frugalité = concevoir juste ?

Explicitement :

La sobriété consiste à :
- réduire les usages,
- modérer la consommation,
- limiter les excès,
- adopter des comportements et des organisations plus économes.

La sobriété : une logique de modération

La frugalité consiste à :
- concevoir autrement dès le départ,
- intégrer les contraintes de ressources dans la solution elle-même,
- chercher le minimum nécessaire,
- éviter d'avoir à "corriger après" un système déjà trop lourd.

La frugalité : une logique de conception sous contrainte

Peut-on être sobre sans être frugal ?

Peut-on être frugal sans être sobre ?

Sobriété
- on part souvent d'un système existant,
- puis on cherche à réduire sa consommation

Frugalité
- on repense le système à la source,
- en intégrant les limites dès le début.

Orange Restricted

Notion d’utilité ?

Une solution numérique est utile si elle :

- répond à un besoin réel
- apporte un bénéfice clair
- concerne des bénéficiaires identifiés
- reste proportionnée
- a un coût global acceptable
- présente un bénéfice net face aux impacts

À questionner systématiquement

- Utile pour quoi ?
- Utile pour qui ?
- Utile par rapport à quoi ?
- Utile jusqu’à quel coût ?

Point d’alerte : Utilité perçue ≠ utilité réelle

Orange Restricted

KPI de sobriété : une boussole pour chaque projet

À suivre systématiquement

Dimension	KPI possibles
Utilité	besoin réel, bénéficiaires, alternative plus simple
Performance	performance suffisante, latence, robustesse
Coûts	TCO, coût d’inférence, coût de maintenance
Environnement	énergie, CO2eq, eau, intensité carbone
Effets indirects	rebond, dépendance, perte de compétences

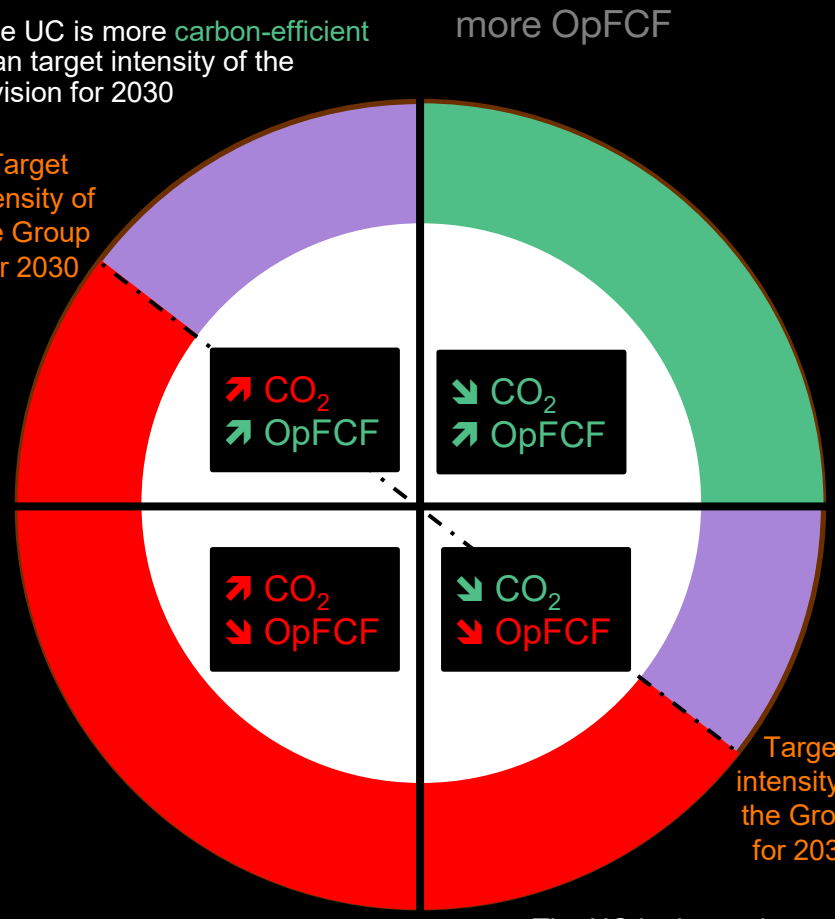
Orange Restricted

How it helps assessing carbon efficiency of AI?

Le '*Carbon Value Navigator*'

https://hellofuture.orange.com/en/the-carbon-impact-of-ai-oranges-navigator-for-the-net-zero-carbon-transition/

The Use-Case increases revenues but may impair the Division's carbon footprint, compared to 2030 target



The carbon efficiency of the service will improve.

Introducing AI will generate more OpFCF as well as decrease carbon

The Use-Case increases Carbon and decreases OpFCF

Bad idea

The Use-Case is net decreasing Carbon but also Operating Free Cash-Flow

Orange Restricted

OpCF = Operating Free Cash-Flow

Effet Rebond - Sans gestion des usages, l'optimisation énergétique et carbone risque d'accentuer les impacts : comment limiter les effets rebonds ?

Si cette efficacité :

- baisse les coûts,
- facilite l'accès,
- accélère les déploiements,
- banalise l'usage,

alors elle peut conduire à :

- plus d'utilisations,
- plus de volumes,
- plus de dépendance,
- et finalement plus d'impact total

On ne peut pas compter uniquement sur l'efficacité technique ; pour éviter les effets rebonds, il faut piloter les usages, cadrer les besoins, fixer des limites et gouverner les déploiements. Autrement dit, rendre une technologie plus efficace ne suffit pas.

Orange Restricted

Effet rebond (illustration)

Indirect Effect	Examples for AI's environmental impacts
Substitution impacts	Art replaced by AI-generated imagery Encyclopedias and books replaced by AI-generated content
Space rebound effects	Datacenters are getting bigger while devices are getting smaller
Scale effects	AI models growing in size and complexity Optimization approaches for AI training and deployment (batching, caching)
Direct economic rebound effects	AI hardware getting more efficient, yet datacenter energy usage is rising
Indirect economic rebound effects	More consumer devices (speakers, appliances, etc.) that allow users to interact with AI systems Smartphones replaced more often to run newest AI functionalities
Economy-wide rebound effects	Power purchase agreements from AI compute providers enable global renewable energy capacity to be expanded
Induction impacts	Targeted ads that use AI to induce more consumption
Time rebound effects	AI-optimized navigation Robots doing household chores

Table 1. Examples of AI's indirect impacts and rebound effects, expanding on Börjesson Rivera et al [95].

From Efficiency Gains to Rebound Effects: The Problem of Jevons' Paradox in AI's Polarized Environmental Debate
Luccioni and al, 2025

Orange Restricted

Le numérique responsable est-il compatible avec les logiques actuelles de croissance des usages ?

Compatibilité partielle, mais tension forte.

Pas vraiment, en tout cas pas spontanément.

Le numérique responsable peut être compatible avec certains développements d'usage, mais pas avec une logique de croissance illimitée, non gouvernée et fondée sur l'effet d'entraînement technologique.

Autrement dit :

- oui si la croissance est sélective, arbitrée, pilotée et orientée vers des gains nets démontrables ;
- non si elle repose sur :
  - l'augmentation continue des volumes,
  - la banalisation des usages,
  - la recherche de performance maximale,
  - et l'absence de limites sur les infrastructures

la croissance des usages n'est jamais "immatérielle". Elle s'appuie sur une base matérielle qui rend la question de la compatibilité très exigeante.

Orange Restricted

Le numérique responsable est-il compatible avec les logiques actuelles de croissance des usages ?

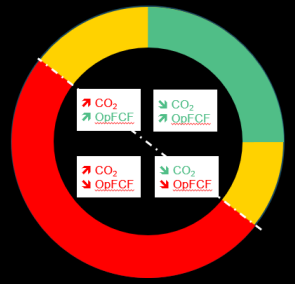
Des logiques actuelles de croissance à questionner :

- **La tyrannie du time-to-market**
- **L'extension des usages par facilité**
- **L'assimilation de l'innovation à la montée en puissance**

Le numérique responsable suppose au contraire une logique de sélection, de proportion et de limite

Il est compatible avec une croissance qualifiée, c'est-à-dire :

- **justifiée par un besoin réel,**
- **proportionnée,**
- **soutenable,**
- **et mesurée en impact global.**



Le '*Carbon Value Navigator*' oppose plusieurs cas :

- **plus de valeur / moins de carbone → favorable**
- **plus de valeur / plus de carbone → à arbitrer**
- **moins de valeur / plus de carbone → mauvais choix**
- **moins de valeur / moins de carbone → à discuter selon contexte**

👉 Donc la croissance des usages n'est compatible avec le numérique responsable que si chaque nouveau cas d'usage améliore suffisamment le ratio valeur / carbone ?

Passer d'une logique de croissance à une logique d'allocation?

Orange Restricted

AFNOR General Reference

A pioneering ADEME + AFNOR initiative to define and promote the notion of “frugal AI”:

- Around sixty companies, institutions and universities involved
- A defining effort around AI and its life cycle
- A catalog of detailed and prioritized best practices via a transparent method
- A desire to take the approach to a European scale

Orange Restricted

Cas d’usage — Classification de verbatim

Illustration du “juste nécessaire technologique”

Objectif métier

Classer automatiquement des verbatim clients / utilisateurs pour :

- identifier des thèmes,
- détecter des irritants,
- mesurer un sentiment,
- prioriser des actions.

Type de tâche

- classification de texte,
- catégorisation,
- analyse de sentiment,
- routage d’insatisfaction ou de motifs de contact.

Ce cas d’usage illustre un principe de frugalité

Avant de choisir la technologie, il faut se demander :

- quel est le niveau de performance réellement nécessaire ?
- faut-il vraiment de l’IA générative ?
- un modèle classique supervisé suffit-il ?
- quel est le coût d’inférence à grande échelle ?
- quelle est l’empreinte carbone du traitement ?
- quel est le risque d’effet de mode ou de suréquipement ?

Pour classer des verbatim, une solution plus récente ou plus “puissante” n’est pas forcément la plus pertinente : un modèle simple peut être plus précis, plus rapide et beaucoup moins énergivore qu’un LLM.

Orange Restricted

Cas d'usage — Classification de verbatim

	Tokens			Embeddings				LLMs			
	Khiops 10multitable Lemmatize	Khiops 10multitable brut	Khiops 11	LGBMClassifier	RandomForestClassifier	LogisticRegression	Khiops 11	gpt-4.1-nano	gpt-4.1-mini	gpt-4.1	claude4-sonnet
total energy cost for 1 inference (Wh) for statistical models: vectorization + transform for LLMs: prompting estimation via ecologits	3,329E-06	3,329E-06	3,329E-06	3,169E-04	3,170E-04	3,168E-04	3,201E-04	5,104E-04	1,117E-03	5,280E-03	5,840E-03
total energy cost for 1 inference (Ref) for statistical models: vectorization + transform for LLMs: prompting estimation via ecologits	1	1	1	100	100	100	100	200	300	1600	1800
Lost luggage recall	71%	70%	72%	91%	90%	86%	75%	87,93%	87,93%	87,93%	87,93%
Average F1 score	66,0%	64,0%	68,8%	83,2%	82,4%	77,2%	75,6%	67,8%	68,1%	68,3%	68,0%
Response time (ms)	1,543103448	1,54310345	1,11706897	36,0332069	36,14393103	36,0171207		385	385	378	378

Merci à Romain Carbou et Fréderic Walch pour les résultats

Sur ce cas de classification de texte :

Enseignement

- le modèle génératif n'est **ni le plus précis**, ni le plus sobre
- une solution classique peut être **meilleure et bien moins coûteuse**
- la bonne question est :

quelle technologie suffit pour répondre au besoin ?

Orange Restricted

La mesure est devenue nécessaire !

La mesure n'est pas accessoire, elle est devenue indispensable pour :

- rendre visible l'impact du numérique et de l'IA,
- comparer des alternatives,
- arbitrer entre performance, coût et carbone,
- justifier des décisions,
- intégrer les impacts dans la gouvernance

- AI Energy Scoring
- CodeCarbon
- ML CO2
- LCA
- TCO / ROI
- Value Clarification Matrix
- AFNOR Spec
- besoin de **benchmarks** et de **standardisation**

Il n'existe pas encore :

- de cadre unifié,
- de métrique pleinement stabilisée,
- de périmètre universellement partagé.
- il faut harmoniser les approches,
- choisir les bons indicateurs,
- aller au-delà du carbone,
- intégrer :
  - eau,
  - équipements,
  - cycle de vie,
  - effets indirects,
  - effets rebond.

Le *Carbon Value Navigator* lui-même précise une limite importante : la méthode est un enabler in the decision process et ne peut pas être utilisée pour une communication externe sur l'impact carbone réel. C'est une remarque très précieuse : elle montre que même un outil de mesure sophistiqué reste un outil d'aide à la décision, pas une vérité absolue. 👉 Cela suggère une vigilance : la mesure peut devenir un instrument de cadrage du récit, pas seulement de transformation réelle.

beaucoup d'impacts restent
invisibles dans le débat public

👉 Cela crée un espace où une organisation peut dire qu'elle mesure, tout en ne couvrant qu'une partie du problème.

Orange Restricted

La mesure est devenue nécessaire !

Une vraie tension : pilotage vs conformité

Deux logiques possibles:

- **Logique 1 — Mesurer pour piloter**
  - comparer, arbitrer, renoncer, améliorer
- **Logique 2 — Mesurer pour afficher**
  - rassurer, se conformer, légitimer, communiquer

l'acculturation un test de sincérité ?

- **rendre les impacts visibles,**
- **former,**
- **diffuser les bonnes pratiques,**
- **développer la compréhension des coûts réels.**

La vraie question n'est pas seulement : "mesure-t-on ?" mais : "la mesure a-t-elle un effet sur les décisions ?"

Mesurer vraiment, c'est accepter que la mesure puisse faire renoncer, ralentir, réduire ou redessiner un projet.

Si la mesure ne peut jamais conduire à :

- **dire non,**
- **réduire le périmètre,**
- **choisir une option moins "prestigieuse",**
- **ou revoir les ambitions à la baisse**

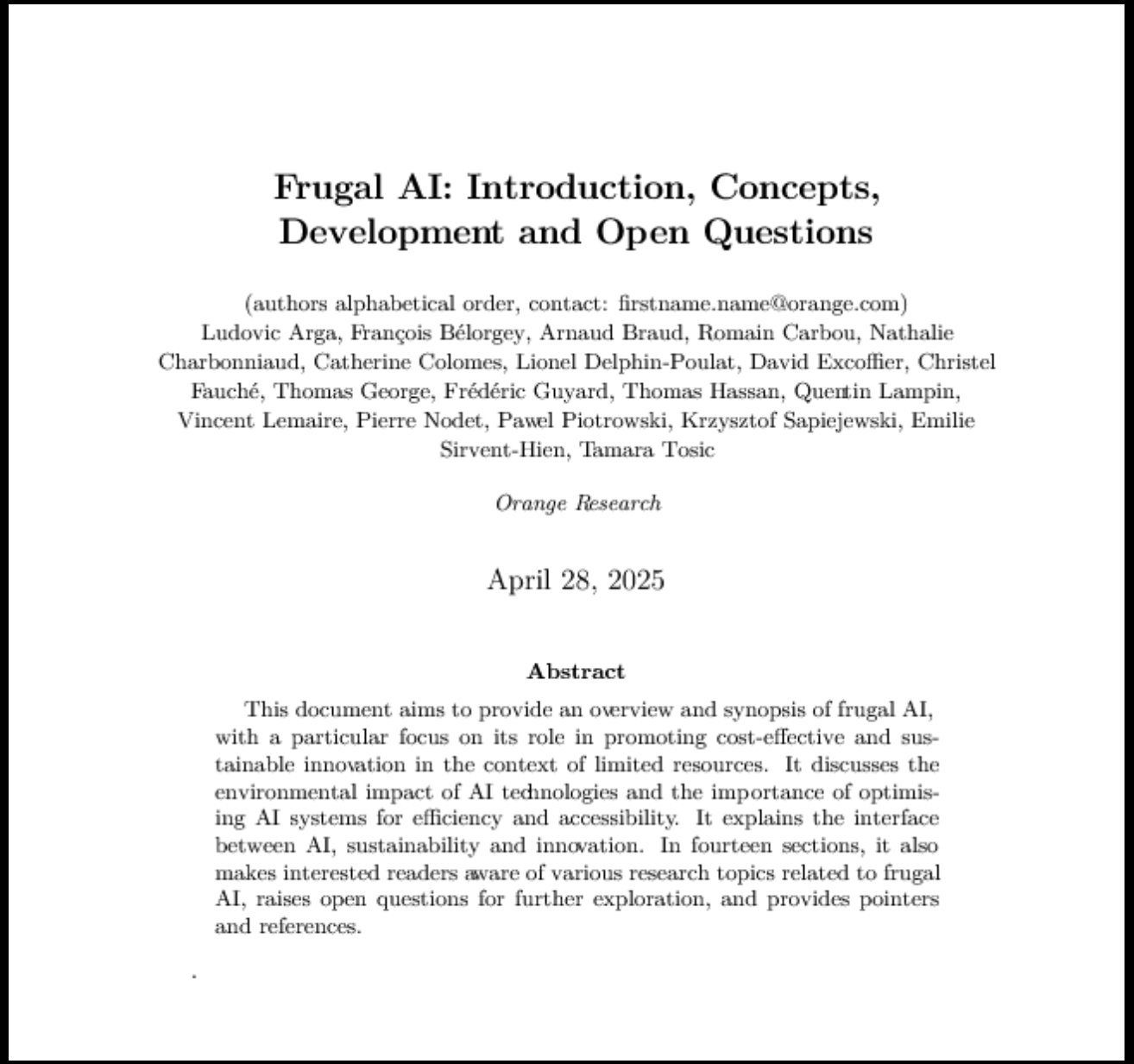
alors elle risque de n'être qu'un outil d'habillage.

Orange Restricted

Frugal AI

Research paper

orange™



Frugal AI: Introduction, Concepts, Development and Open Questions

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

Orange Research

April 28, 2025

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.

Orange Restricted

Articles ‘grand public’ associés

Frugal Artificial Intelligence: Why, what and how?

AI growth and global electricity production: incompatible?

Use the right AI for the right need at the right time

AI’s environmental footprint: understanding, measuring, acting

Orange Restricted

Petit TP

Orange Restricted

Thanks

Orange Restricted