Frugal Artificial Intelligence: Challenges and Approaches

NGUYEN Kim Thang

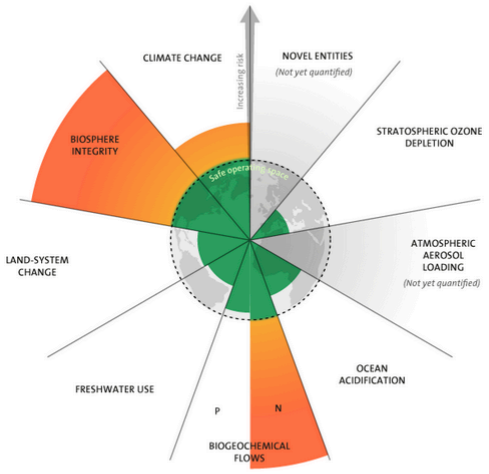
Professor at Ensimag,
Grenoble INP, Université Grenoble Alpes
Director of Education,
Multidisciplinary Institute in Artificial Intelligence (MIAI Cluster)

Observations and Trends

Current trends

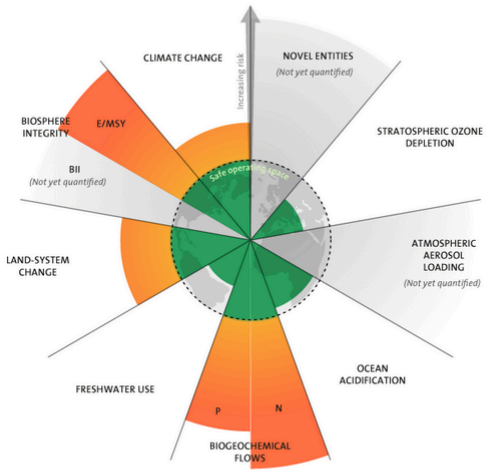
- Global warming
- A major environmental crisis: visible signs
- The Earth system is becoming destabilized

2009



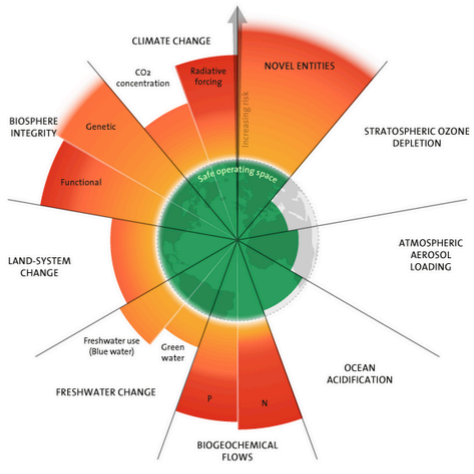
7 boundaries assessed,
3 crossed

2015



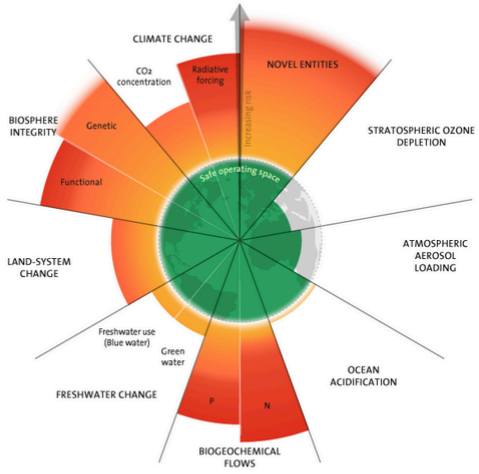
7 boundaries assessed,
4 crossed

2023



9 boundaries assessed,
6 crossed

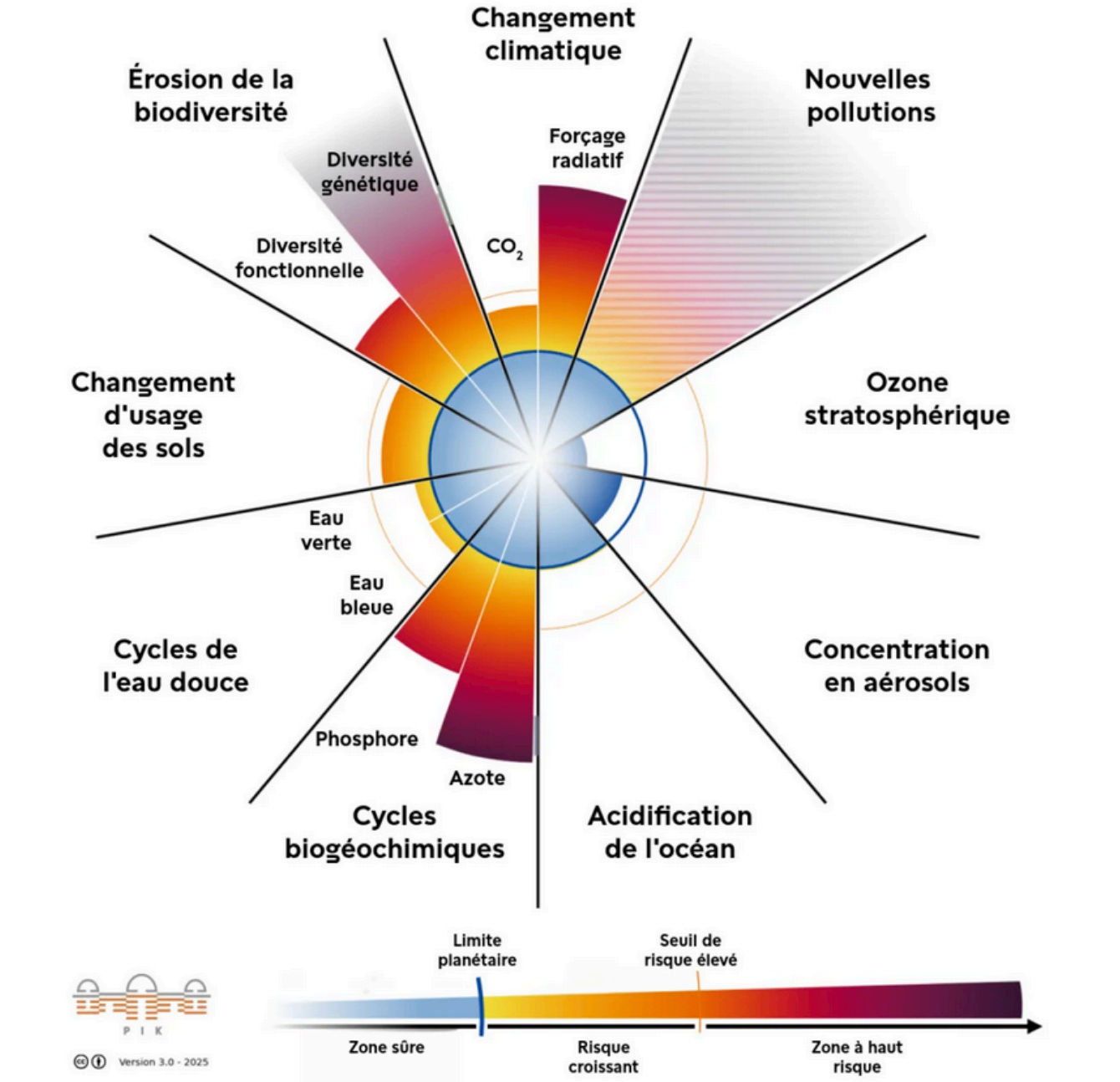
2025



9 boundaries assessed,
7 crossed

Observations and Trends

- Human activities: a major cause
- Public awareness



État des limites planétaires en 2025 CC-BY PIK, traduit par MNHN

The Digital Sector's Share

Ecosystem: An ecosystem encompassing digital devices and infrastructure (servers, end-user devices, networks) and the digitalization of society (industry, transport, healthcare, etc.)

Connected devices

- 2024: 18.8 billion, an increase of 13%
- Projection for 2030: 40 billion

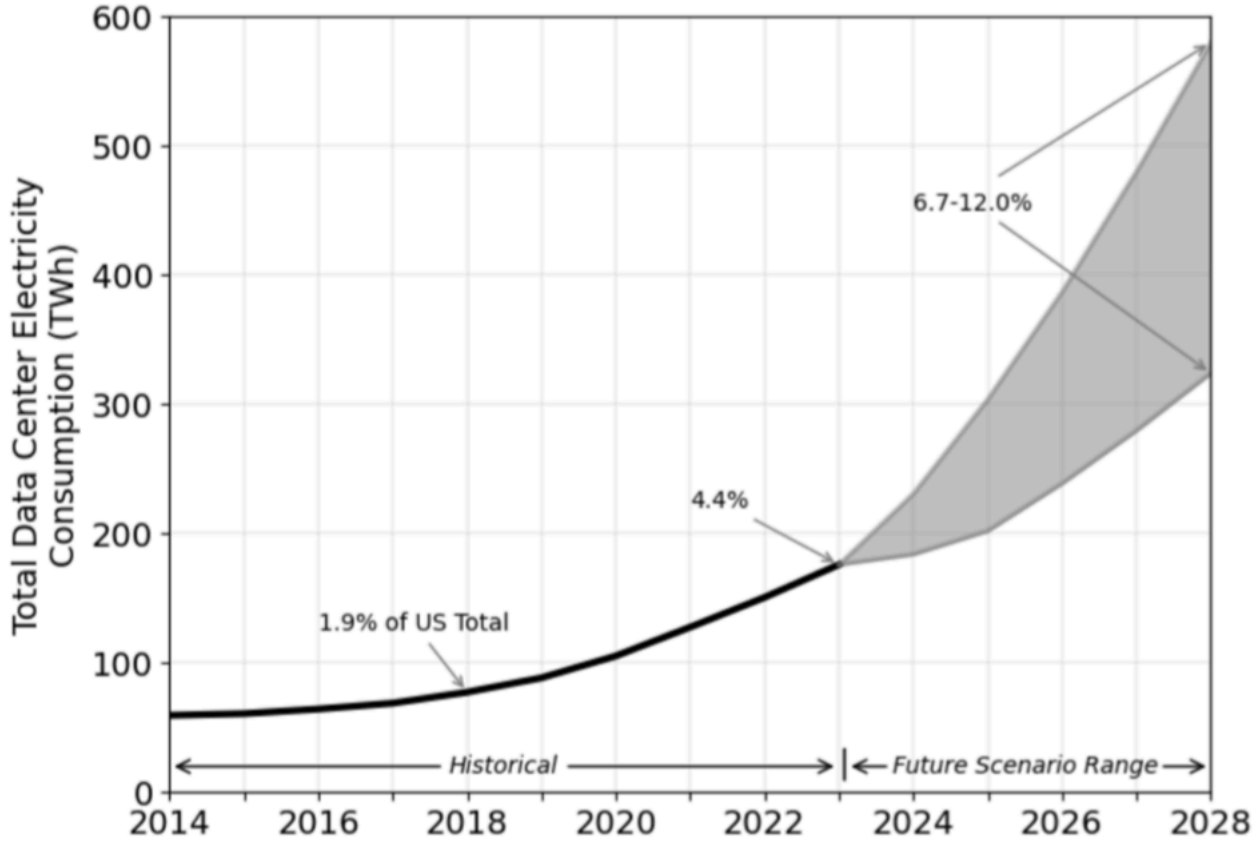
Computing platforms

- 1985: Gigaflops/second (10^9)
  - 200 W/flops
- 1997: Teraflops/second (10^{12})
  - 0.85 W/flops
- 2008: Petaflops/second (10^{15})
  - 0.002483 W/flops
- 2024: Exaflops/second (10^{18})
  - 0.000029581 W/flops

The Digital Sector's Share

Data centers

- 4.4% of US electricity consumption
- 1.5% of global electricity consumption
- Projection for 2030: twice as much, equivalent to Japan's electricity consumption



The digital sector

- In France: 10% of electricity consumption
- 4% of total carbon emissions
- AI: an accelerating factor and one of the main drivers of growth

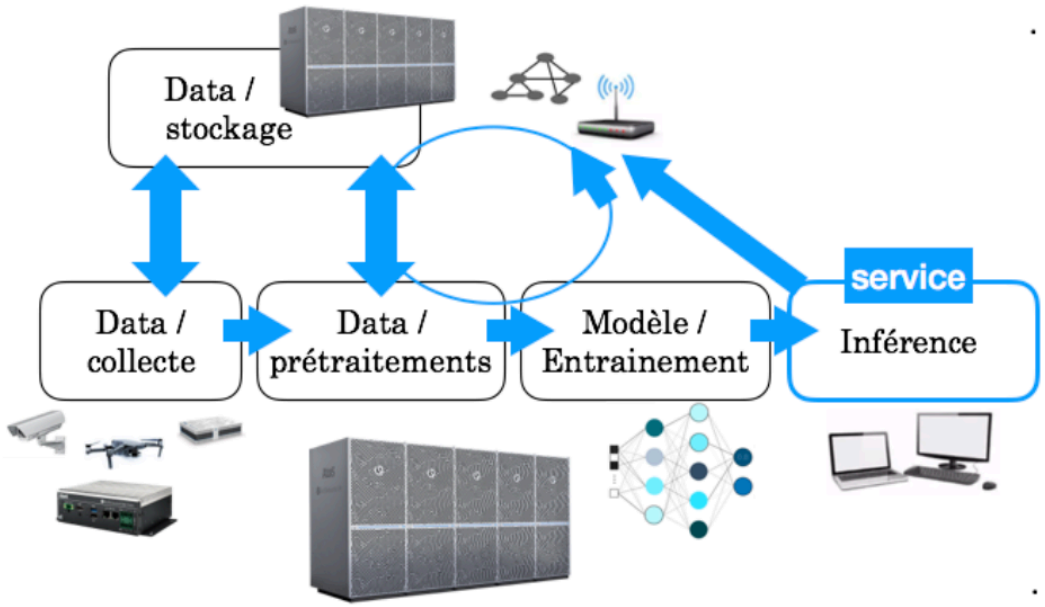
AI and Generative AI

Key milestones in AI

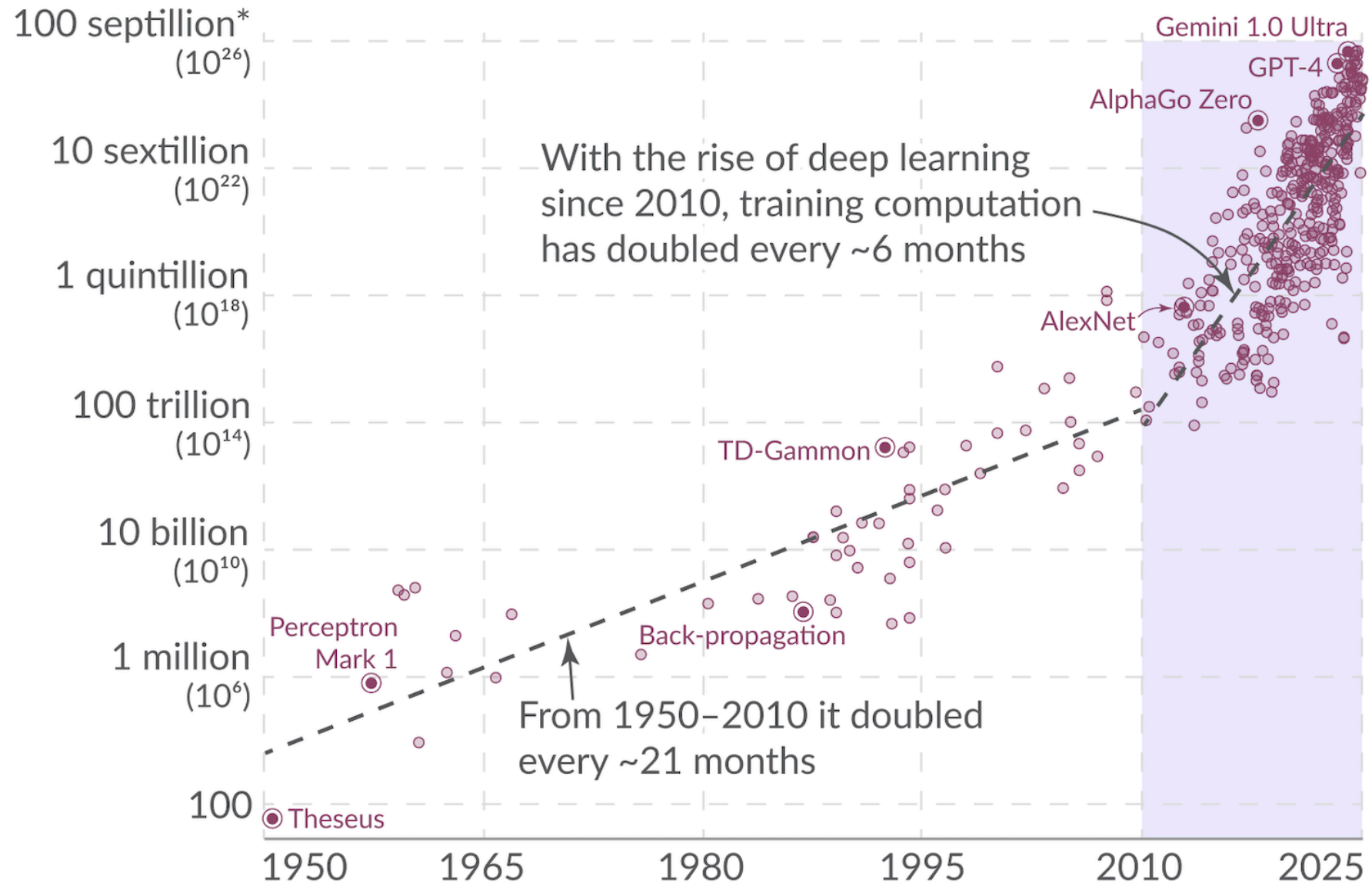
- 2013–2014: Deep learning (computer vision)
- 2016: Reinforcement learning (AlphaGo)
- 2018: Transformers
- 2022: Generative AI (ChatGPT)

AI systems

- Model training
  - Training / validation / testing / production
- Inference: using trained models



AI and Generative AI



source: Our World In Data

Approaches to Frugal AI

Technical approaches

- **Optimization strategies**: minimize environmental impacts while maintaining reasonable performance
- **Measure** and **estimate** impacts
- **Design** new models and architectures; reduce complexity, model size, and data requirements
- **Low-energy computing**: federated learning, embedded AI, computing facilities and their energy mix
- **Resilience**: reduce dependence on infrastructure and extend equipment lifespans

Systemic approaches

- Study **direct** and **indirect effects**
- Raise **awareness** of usage practices / **regulation**
  - AFNOR guidelines: a framework for assessing and reducing environmental impact

Measurement and Assessment

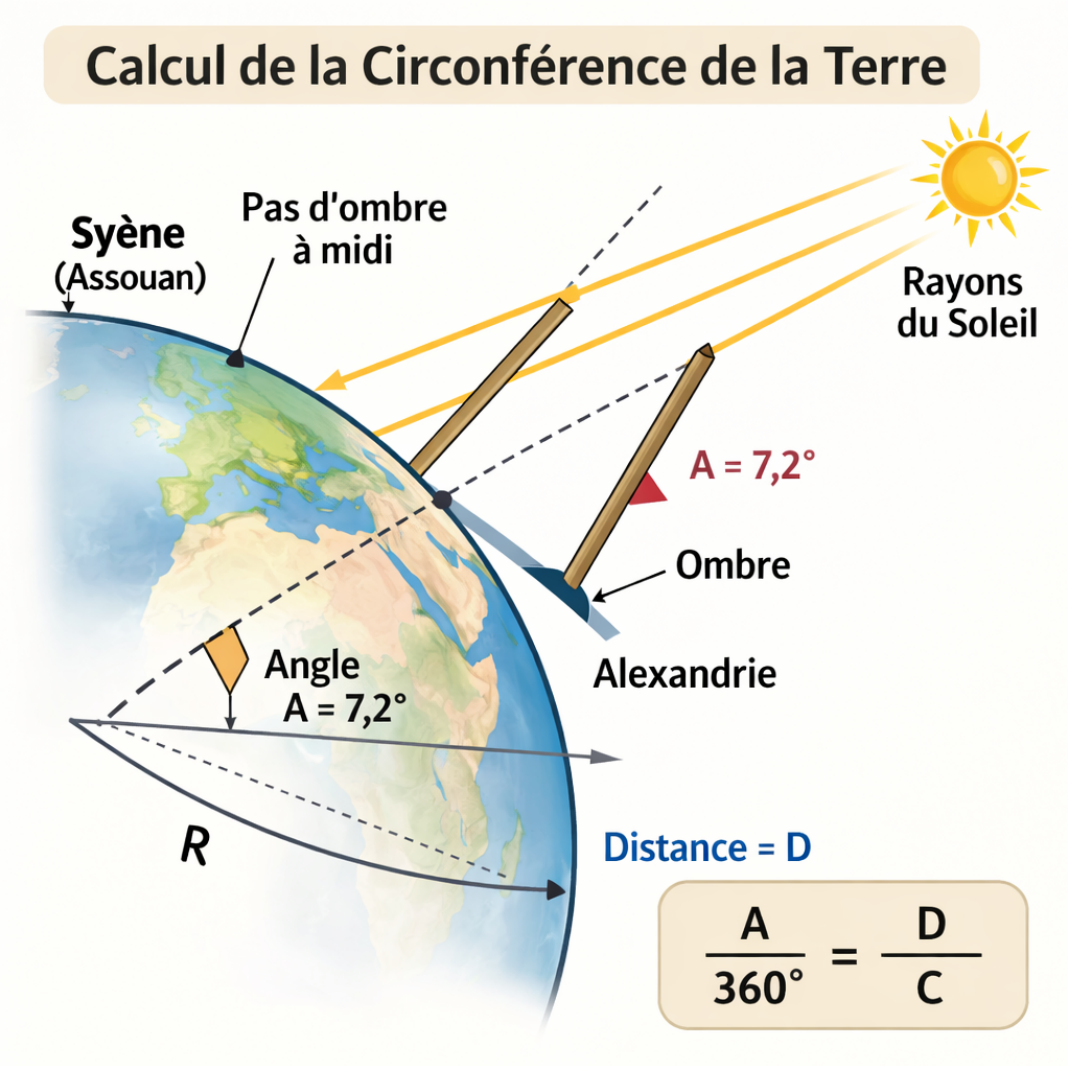
Without measurement, there is no science

- Quantify
- Make assessments objective
- Put results into perspective to enable comparisons

Good measurement

- A clearly defined object of measurement
- Methods:
  - Agreed upon and practical
  - Acceptable within the context of its time

Historical Example: The Earth's Circumference



- Distance between Syene and Alexandria: 5,000 stadia
- Angle measured in Alexandria: 7.2 degrees, or 1/50 of a circle
- Circumference: 50 × 5,000 stadia
  - Equivalent to 39,375 km
  - Modern value: 40,075.017 km
  - Impressive: an error of less than 2%
- Unit of measurement: the length of a stadion
- High uncertainty
- Assumptions:
  - The Earth is perfectly spherical
  - Alexandria and Syene lie on the same meridian
  - Small-angle approximation

Measuring Environmental Impact

Google Gemini

MEASURING THE ENVIRONMENTAL IMPACT OF DELIVERING AI AT GOOGLE SCALE

Cooper Elsworth, Keguo Huang, David Patterson, Ian Schneider, Robert Sedivy, Savannah Goodman, Ben Townsend, Parthasarathy Ranganathan, Jeff Dean, Amin Vahdat, Ben Gomes, and James Manyika

Google, Mountain View, CA, USA

Results

- 0.24 Wh per median prompt
- 0.26 mL of water consumed — approximately 5 drops

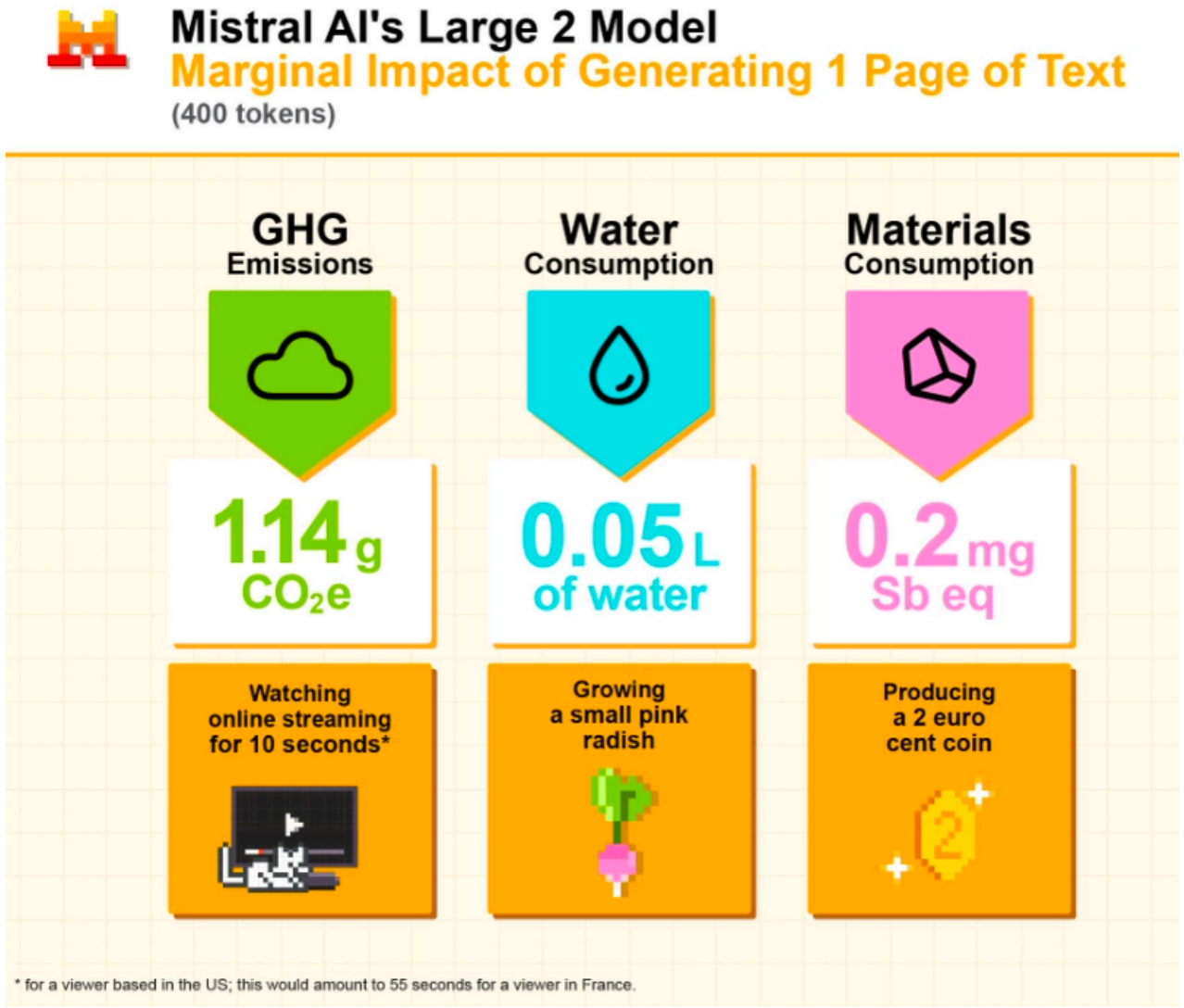
Comments

- Partial transparency: no open data, and methods only partially available
- Unclear functional unit: “median prompt”
- Favorable energy accounting through power purchase agreements

Measuring Environmental Impact

Mistral

- A more clearly defined functional unit
- A more robust methodology: AFNOR specification for frugal AI
- …but no accompanying data



Measurement and Assessment in AI

Without measurement, there is no science

- Quantify
- Make assessments objective
- Put results into perspective to enable comparisons

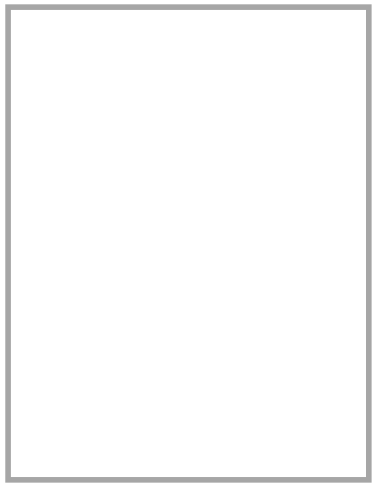
Good measurement

- A clearly defined object of measurement
- Methods:
  - Agreed upon and practical
  - Acceptable within the context of its time

Measurement in AI

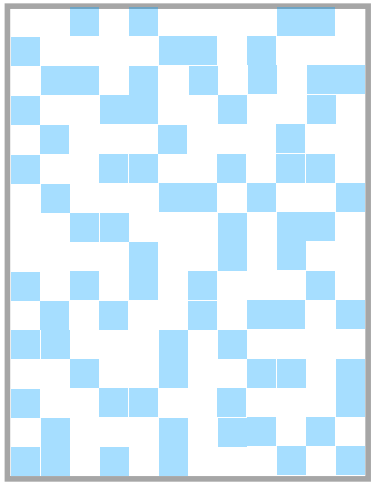
- A long road ahead…
- Measurements alone will never be sufficient
- Beyond efficiency: direct and indirect effects

Rebound Effects



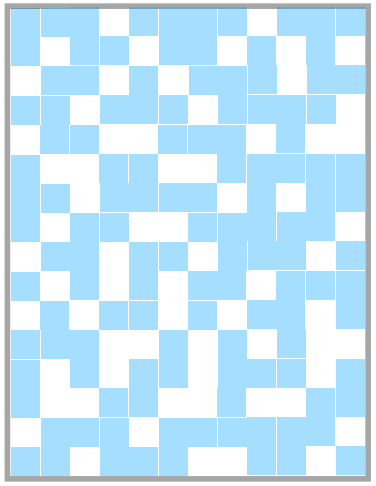
An available resource, such as electrical power in a data center.

Rebound Effects

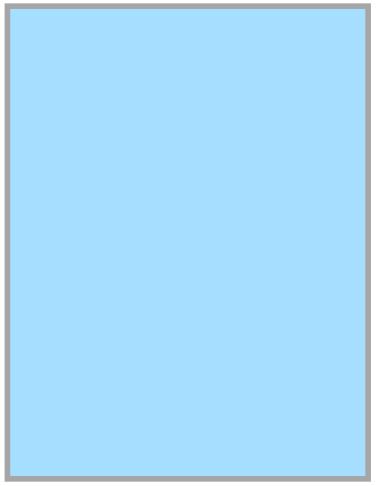


- Consumption of one unit of computation.

Rebound Effects

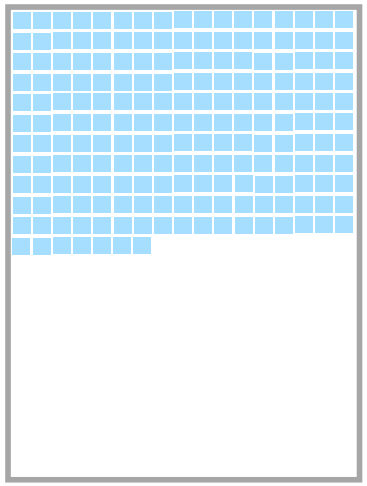


Rebound Effects



The total available resource is fully used by demand for computation.

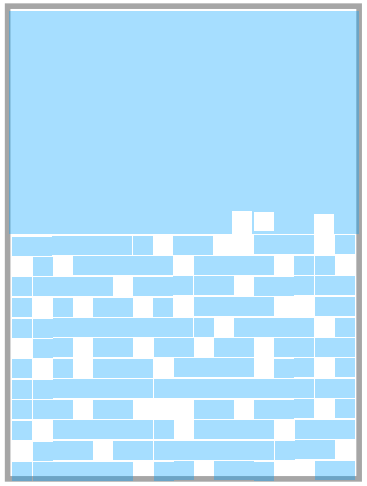
Rebound Effects



Technological improvement: Each computation requires less energy

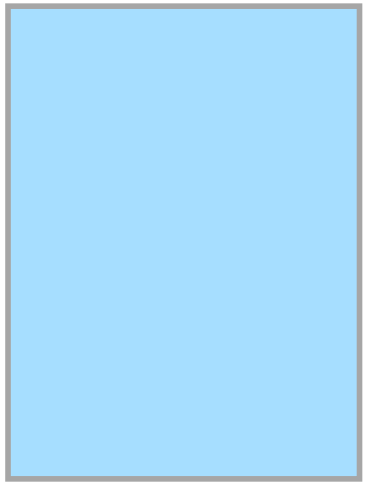


Rebound Effects



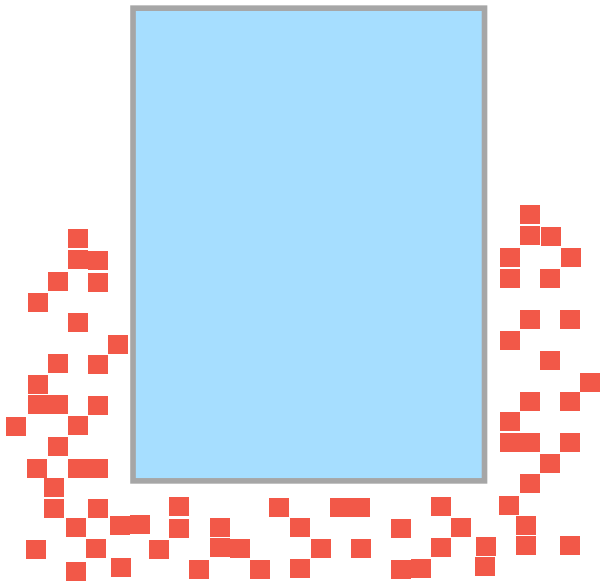
Additional computations can now be performed.

Rebound Effects



The available electrical power is once again fully used.

Rebound Effects

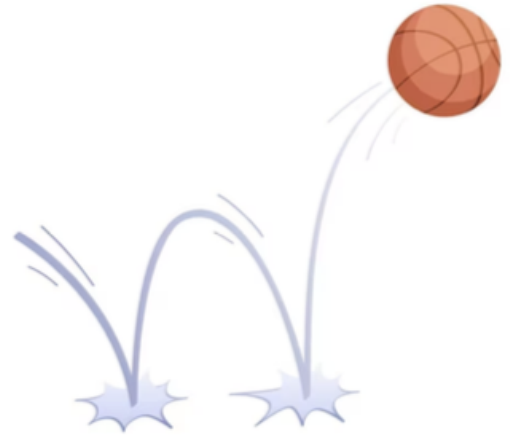


This is the rebound effect.

Rebound Effects

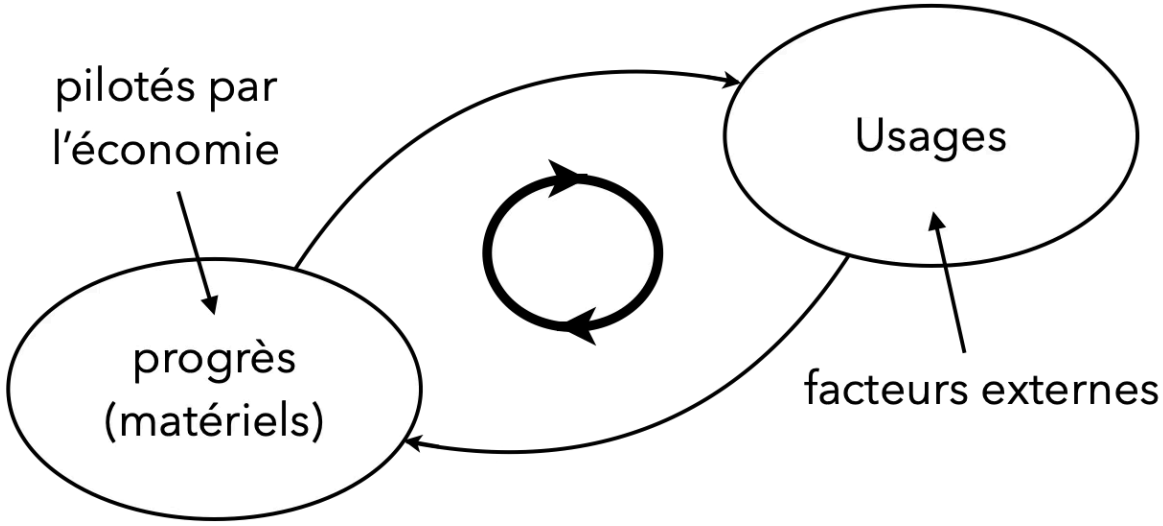
Direct rebound effects

- A technological advance improves a technical system, for example in its use of energy resources
- Usage increases until the resource is fully used
- A new solution allows the same amount of work to be done with fewer resources
- This progress enables further expansion



Indirect rebound effects

- Indirect rebound also occurs when efficiency gains lead to other uses



Rebound Effects: A Paradox

Information

- Impacts cannot easily be predicted until the technology has been widely developed and used

Power to act

- Controlling or changing a technology becomes very difficult once it is firmly established

Designing technological tools

- The Collingridge dilemma
- Interdisciplinary approaches that go beyond efficiency and performance

Designing Convivial AI Tools

Principles

- Question the usefulness of the intended service
- Ensure a small environmental footprint compatible with limited resources
- Resilience: the ability to adapt to instability, operate with reduced functionality, and be repaired
- Plan from the outset to extend the service's lifespan
- Acceptance: ensure fair deployment through consultation across the relevant context and stakeholders

Frugal Artificial Intelligence

Frugal artificial intelligence is an opportunity

- Ecological transition
- More than an ecological necessity: a competitive advantage
- Aligned with trustworthy AI and explainable AI

Conviviality in AI system design

- Usefulness: necessity / need
- Environmental impact
- Resilience
- Durability
- Sustainability
- Control
- Equity / acceptability