The Supposed Virtuality of AI Meets the Wall of Reality

How USD 200 Billion in GPUs Reveals the Physical Constraints of Artificial Intelligence

Jérôme Vetillard

VP R&D | Chief Product Officer | AI-Powered Healthcare & Life Sciences Products

Compliance by Design | PhD AgroParisTech | MIT Sloan | IE Business School & Brown University

Revised edition, February 2026

Executive Summary

Between 2022 and 2025, hyperscalers invested more than USD 200 billion in GPU infrastructure for generative AI, a rational real-options strategy in the face of radical uncertainty over the size of the market. Two years later, enterprise adoption remains marginal (less than 10 percent of U.S. firms, 74 percent extracting no tangible value), creating an imbalance between installed capacity and effective usage.

To absorb this overcapacity, the industry is pushing toward intensive uses such as persistent autonomous agents, shifting the bottleneck from GPUs toward energy, water, and training data. AI is entering a heavy-industry regime in which competitive advantage is played out on energy contracts and water rights as much as on algorithmic excellence. The ultimate question is no longer technical but political: what share of our finite physical resources will our societies agree to allocate to AI, and at the expense of which other uses?

I. Introduction: the paradox of constrained abundance

We speak of artificial intelligence as if it were an immaterial technology: algorithms, models, tokens generated in the abstraction of the cloud. This representation dominates strategic discourse, market analyses, and adoption forecasts. Yet it is becoming increasingly misleading.

Behind every query to a language model, every generated image, every line of assisted code, lies a physical reality: tens of thousands of GPUs aligned in datacenters, electricity consumption comparable to that of a medium-sized city, significant volumes of water for cooling, and connections to power grids already under strain. Generative AI is no longer merely an “augmented software” industry: structurally, it is trending toward heavy industry.

Between 2022 and 2025, the hyperscalers (Microsoft, Google, Meta, Amazon) engaged in an unprecedented investment race, primarily in compute infrastructure. This accumulation was neither irrational nor accidental: faced with radical uncertainty about uses and the scale of the market, securing capacity ahead of competitors was a defensible strategic option. The cost of temporary excess appeared lower than the risk of being capacity-constrained at the moment the market tipped.

Two years on, a gap is becoming clearer: enterprise usage is progressing, but remains far from a massive and homogeneous deployment. Returns on investment are diffuse, and usage is still frequently sporadic, centred on non-persistent human interactions. The initially assumed overcapacity then becomes an economic imbalance: a very large installed infrastructure, imperfectly monetised.

This tension is what is now pushing the industry toward more intensive uses: persistent agents, multi-model orchestrations, continuous workloads. These paradigms dramatically increase the utilisation rate of machines, but they shift the binding constraint. The bottleneck is no longer primarily the number of available GPUs. It tends to become the capacity to durably mobilise gigawatts that can be connected to the grid, volumes of water, high-quality training data, permits, and sufficient social acceptability.

This shift redefines the rules of competition: advantage slides away from laboratories alone, toward long-term energy contracts, infrastructure engineering, locational choices, and territorial trade-offs. AI is entering a regime in which its technical possibilities collide with thermodynamic constraints and with the governance of finite physical resources.

This transformation is not a mere growing pain. It is a change of industrial regime. It raises a question we have not yet confronted head-on: will artificial intelligence be limited by our algorithmic creativity, or by the thermodynamic and political constraints of the planet that supports it?

II. 2022–2024: pre-emptive accumulation

When ChatGPT crossed the one-million-user threshold in five days, in November 2022, nothing was yet stabilised: not the ultimate use cases, not the speed of adoption, not the real size of the market. Generative AI could become a niche tool or a universal platform. This radical uncertainty might have led to wait-and-see. It produced the opposite: an unprecedented mobilisation of capital.

Between 2022 and 2025, hyperscalers collectively committed more than USD 200 billion to compute infrastructure. Microsoft announced USD 50 billion over four years, Meta USD 37 billion for the single year 2024, and Amazon USD 75 billion dedicated to AI datacenters. These amounts far exceed the investments observed during earlier technology waves (cloud, mobile, big data).

This accumulation follows a well-identified economic logic: that of real options. Under radical uncertainty, building excess capacity amounts to purchasing a strategic option. The cost of this option (temporarily under-utilised GPUs) is judged to be lower than the risk of being excluded from an emerging market with increasing returns, for lack of available capacity at the critical moment. Microsoft is not building for the demand of 2023, but for the hypothetical and massive demand of 2026–2027. The overcapacity is not suffered: it is assumed.

This strategy immediately collides with a physical constraint: silicon. Producing an NVIDIA H100 GPU requires etching processes at 4–5 nm, mastered almost exclusively by TSMC. Lead times reach 18 to 24 months, rendering supply structurally inelastic in the short term. GPUs then become strategic

assets. Contracts with NVIDIA involve prepayments, volume commitments, and, in some cases, allocation priorities. Microsoft secures access to several hundred thousand H100s, Meta develops its own chips (MTIA), Amazon deploys Trainium and Inferentia, and Google has been operating its TPUs since 2016.

In 2023–2024, GPUs become the object of a veritable arms race. Their value is no longer merely functional, but positional. China, sidelined by American protectionism, launches its own "Manhattan Project" to master etching below 5 nm.

This dynamic produces an extreme concentration of compute resources. At the end of 2024, Microsoft reportedly targets about 1.8 million GPUs, while Stanford University has only around 300 GPUs (Russell Wald, Stanford HAI). Even ambitious academic initiatives remain marginal: the Marlowe superpod at Stanford (248 H100s), the Kempner Institute cluster at Harvard (384 H100s), and the 32 H100s available at MIT in 2024, all sit three to four orders of magnitude below the capacity of a single hyperscaler. At a U.S. Senate hearing in November 2024, Russell Wald summarised the situation plainly: "All U.S. universities combined could not build a version of ChatGPT today."

Frontier publications are progressively migrating from university laboratories to the in-house teams of hyperscalers. AI is becoming an experimental science reserved for actors able to commit several billion dollars in CapEx. This pre-emptive accumulation fits squarely within the theory of increasing-returns markets (Arthur, 1989): network effects, high switching costs, economies of scale, and winner-takes-most dynamics. Satya Nadella put it explicitly in 2023: "We're not going to be caught short on compute."

As this GPU capacity is deployed, however, a new constraint appears: memory. Modern architectures are now memory-bound rather than compute-bound. An H100 carries 80 GB of HBM3; a GB200 promises 192 GB, but next-generation models require several terabytes of aggregated memory. The HBM industry, dominated by SK Hynix, Samsung, and Micron, struggles to keep up: 8-to-12-month lead times, and price increases estimated at +300 percent between 2022 and 2024.

At the end of 2024, the industry holds several million advanced GPUs, amounting to an aggregate compute power above 100 exaflops, well beyond immediate measurable usage. This is precisely the outcome sought ex ante. But this ex-ante strategic rationality will collide, ex post, with a more complex reality.

III. 2024–2025: the under-consumption trap

3.1. The adoption gap: converging data

Two years after the launch of ChatGPT, a structural gap has emerged between installed compute capacity and effective usage. The data converge: enterprise adoption of generative AI remains marginal, fragmented, and largely unprofitable.

According to the U.S. Census Bureau, only 9.2 percent of U.S. firms declared using AI in the second quarter of 2025, compared with 5.7 percent at the end of 2024: a real progression, but well below initial projections. A BCG study (2024) covering 1,000 executives shows that 74 percent of firms extract no tangible value from their AI initiatives. McKinsey confirms: more than 80 percent of organisations report no material impact of generative AI on their financial results. Only 6 percent of respondents, whom McKinsey labels "high performers," claim an AI-attributable effect above 5 percent on EBIT. This concentration warrants caution: these organisations are precisely those that have invested most heavily, and would be most exposed to a stock-market devaluation in the event of negative announcements on their AI returns. Declarative bias cannot be ruled out.

The problem is not merely deployment, but actual usage. An MIT study indicates that 95 percent of generative AI pilot projects fail to produce measurable returns, not for technical reasons, but for lack of operational scale-up. Usage remains sporadic: where projections anticipated more than 100 monthly queries per active user, the observed reality lies between 8 and 12. Even widely distributed copilots (Microsoft 365 Copilot, Google Workspace AI) struggle to convert perceived utility into measurable economic gains.

More revealing still: certain organisations, Microsoft itself among them, have incorporated Copilot adoption into the performance evaluation metrics of their employees. Indicators such as the "Copilot engagement rate" or "Copilot actions per user" now figure among the KPIs tracked by managers. This practice, sometimes described as "institutional dogfooding," reveals the gap between the narrative of organic adoption and the reality of diffusion by hierarchical incentive. Coerced usage guarantees neither genuine appropriation nor the generation of value.

3.2. Anatomy of the frictions

This under-performance is in no way mysterious. McKinsey shows that close to 70 percent of obstacles are human and organisational, 20 percent are technological, and only 10 percent are algorithmic, even though the latter capture the bulk of attention and resources.

Technical frictions (20 percent) stem from integration with legacy systems and from data quality, which is often fragmented and poorly governed. Generative AI cannot be dissociated from the maturity of data governance; likewise, any MLOps ambition presupposes a pre-established DevOps culture. According to Curt Jacobsen of McKinsey, 30 to 50 percent of innovation teams' time is spent securing regulatory compliance or waiting for organisational policies to evolve.

Organisational frictions (70 percent) run deeper. Resistance to change is not irrational: it reflects uncertainty over the future impact on jobs, the absence of adequate training, and a deficit of strategic vision. A Gallup survey (2024) reveals that only 15 percent of U.S. employees report that their company has communicated a clear AI strategy. Fewer than 30 percent of CEOs directly sponsor the AI agenda, according to McKinsey. Legal frictions intensify as regulation tightens: GDPR, the AI Act adopted in 2024, and questions of liability and explainability. Economic frictions appear ex post: the

real TCO of an AI in production (inference, fine-tuning, monitoring, compliance) turns out to be 10 to 20 times higher than the costs of initial PoCs.

3.3. Demand-stimulation attempts and the agentic pivot

Faced with this under-consumption, the industry tries to activate demand through several levers: a proliferation of marketing use cases meant to demonstrate the universality of the technology; aggressive bundling (Microsoft embeds Copilot across all its enterprise suites, Google does the same with Workspace AI); a price war (OpenAI cuts the price of GPT-3.5 by more than 90 percent between 2023 and 2024); and a massive pivot toward agentic AI from the fourth quarter of 2024.

This last pivot is not accidental. As McKinsey puts it explicitly in its June 2025 report: horizontal copilots have not generated value at scale; actors are turning to autonomous agents, embedded in vertical business processes, able to operate continuously. Yet a question imposes itself: why would organisations unable to make simple uses of generative AI profitable suddenly extract value from systems that are markedly more complex to deploy, govern, and secure?

3.4. Structural economic imbalance and financial risk

The result is a major economic imbalance. Millions of GPUs are installed, but only a fraction is effectively used productively. Real utilisation rates of GPU clusters sit between 15 and 30 percent according to sector analyses, well below the 60 to 80 percent required for such investments to be profitable.

This imbalance raises a financial question that the industry still avoids confronting head-on: do the USD 200 billion invested constitute a rational strategic allocation, or the early stages of a speculative bubble? Several indicators call for vigilance. The ratio between engaged CapEx and AI-attributable revenues remains unfavourable: for every dollar invested in infrastructure, AI-specific revenues generate only a fraction of a return, with the remainder booked under general cloud revenues. The duration of GPU assets is problematic: the cycle of technological obsolescence (18 to 24 months between generations) is significantly shorter than accounting amortisation (5 to 7 years), creating a risk of accelerated depreciation. Finally, the risk of architectural rupture is not negligible: if alternative architectures (Mamba-style SSMs, text diffusion models, hybrid architectures) were to significantly reduce the need for attention-optimised GPUs, entire waves of hardware would become stranded assets.

The most illuminating precedent is the fibre-optic overcapacity of 2000–2001. In the late 1990s, telecom operators had invested massively in fibre networks, convinced that bandwidth demand would grow exponentially. Demand did grow, but with a temporal lag of several years relative to projections. The most exposed companies (WorldCom, Global Crossing, 360networks) went bankrupt. The infrastructure was eventually used, but by other actors, under other conditions, after a massive destruction of shareholder value. The analogy is not perfect, as today's hyperscalers are financially

more resilient than the telecoms of 2000, but the structural pattern (pre-emptive overcapacity, adoption lag, pressure on valuations) is analogous and deserves attention.

This pressure is pushing hyperscalers toward a race for volume: price cuts, bundling, and the invention of new, more intensive usage paradigms. Autonomous agents operating 24/7 consume an estimated 50 to 100 times more resources than a one-off query (an estimate based on the ratio of duty cycles: a persistent agent maintains a memory context and executes inference loops continuously for hours or days, versus a few seconds for an interactive query; the ratio mainly reflects the difference in GPU and HBM memory occupancy time, not a proportional increase in raw compute per unit inference). By seeking to monetise compute overcapacity through the intensification of usage, the industry mechanically shifts the bottleneck.

IV. 2025–2027: the shift toward physical constraints

The intensification of AI uses is no longer primarily constrained by GPU availability, but by the capacity to sustain their continuous operation. The binding factor is shifting toward physical inputs whose expansion is slow, costly, and politically constrained. This shift marks a regime change.

4.1. Energy: the gigawatt wall

Orders of magnitude are enough to gauge the scale of the phenomenon. A hyperscale datacenter typically draws 150 to 300 MW under continuous load, equivalent to the consumption of a city of 100,000 to 200,000 inhabitants. A cluster of 10,000 H100 GPUs requires roughly 20 MW of permanent power, not including cooling.

The energy requirements of training large models remain opaque, but estimates converge around dozens of megawatts mobilised over several months, representing volumes of the order of 100 GWh for GPT-4-generation models. According to the IEA, datacenters consumed about 460 TWh in 2022. Projections to 2030 lie between 1,000 and 1,300 TWh, of which 40 to 50 percent are attributable to AI.

This demand is geographically concentrated, creating severe local tensions. In Ireland, datacenters represented about 20 percent of national electricity consumption in 2023, with projections close to 30 percent by 2030. In Northern Virginia, capacities are approaching saturation. Singapore and the Netherlands have imposed moratoria on new projects. Adding electrical capacity takes time: 3 to 5 years for a gas plant, 10 to 15 years for traditional nuclear.

Faced with these constraints, the industry is turning to nuclear power. Microsoft has signed a long-term agreement with Constellation Energy for the restart of Three Mile Island Unit 1 (approximately 835 MW). A race for SMRs (Small Modular Reactors) is under way: AWS targets more than 5 GW by 2039, Google about 500 MW via Kairos Power, Oracle is designing SMR-powered datacenters, and TerraPower (Bill Gates) is developing the Natrium reactor (345 MW), a fast-neutron reactor using

molten sodium as primary coolant, in the lineage of the French Phénix, Superphénix, and Astrid programmes.

SMRs offer theoretical advantages (modularity, shorter lead times, proximity to load), but introduce new systemic risks. Many rely on HALEU uranium (up to 19.75 percent), whose enrichment rate is close to the military threshold. The potential dissemination of hundreds of reactors complicates international oversight and weakens existing non-proliferation regimes.

4.2. Water: the invisible constraint

Water constitutes the other critical bottleneck, often absent from public discourse. Depending on size, climate, and cooling technology, a hyperscale datacenter may consume several million litres of water per day. Microsoft reported a 34 percent increase in its water use between 2021 and 2022, Google a 20 percent increase, both largely attributed to the expansion of their AI infrastructures.

According to the World Resources Institute, about 40 percent of the world's datacenters are located in zones of medium-to-high water stress. This constraint is geographical and seasonal, fuelling use-conflicts already visible in Arizona, Uruguay, and Spain. Technological alternatives (air cooling, immersion cooling, nordic location) entail severe trade-offs in terms of cost, energy, or latency.

4.3. Training data: the fourth wall

Beyond energy and water, a material constraint emerges that follows the same logic of finitude: the exhaustion of high-quality training data. The scaling laws that govern the improvement of foundation models rest on three inputs: compute, parameters, and data. The first two are extensible through investment; the third proves to be finite.

Villalobos et al. (2022) estimate that stocks of high-quality text available on the public web could be exhausted between 2026 and 2032, depending on assumptions about model growth. Muennighoff et al. (2023) reach convergent conclusions and show that repeated reuse of the same data (multiple epochs) produces diminishing returns beyond a relatively low threshold.

Substitution by synthetic data, generated by the models themselves, raises a fundamental problem. Shumailov et al. (2023) demonstrate that recursive training on synthetic data causes progressive model degeneration ("model collapse"): distributions contract, tails disappear, and the model converges toward an impoverished representation of the world. The phenomenon is analogous to genetic inbreeding: each generation loses informational diversity.

This constraint reinforces the central argument of this article: AI collides not with a single physical limit but with a converging bundle of material constraints, namely energy, water, silicon, and data, which frame its growth on every side. The underlying reason is the same in each case: these resources are finite, their expansion is slow, and their substitution introduces severe trade-offs.

4.4. Geographic lock-in

AI is becoming geographically determined by the convergence of these physical constraints: abundant and dispatchable low-carbon energy, durable hydrological basins, and connectivity to internet backbones with acceptable latency. These conditions are rarely met simultaneously, creating a durable territorial lock-in. Some regions enjoy relative advantages (Nordic countries, Quebec, New Zealand, France), but they remain partial. In France, heatwave episodes and low river flows have already led to reductions in nuclear output, revealing the cross-dependence between energy and water.

A technology reputed to be immaterial, AI is becoming one of the most materially constrained industries of the twenty-first century.

V. Systemic dynamics: the rebound effect at industrial scale

5.1. Jevons' paradox revisited

In 1865, William Stanley Jevons observed in *The Coal Question* that improvements in steam-engine efficiency had not reduced total coal consumption. By lowering the energy cost per unit of work, efficiency had broadened the field of possible uses, producing a net increase in aggregate demand. This mechanism, known as backfire in the rebound-effect literature, only fully manifests in contexts where demand is not saturated and is highly sensitive to marginal costs.

Generative AI meets precisely these conditions. Algorithmic and hardware progress have dramatically reduced the unit cost of compute. GPT-4-generation models produce a token for a fraction (often estimated around one-tenth) of the energy required by GPT-3. Recent generations of specialised processors show performance-per-watt gains on the order of ×2 to ×3 compared with prior generations.

These gains are real. And yet the total energy consumption of the AI ecosystem is growing rapidly. The reason is structural: each efficiency improvement expands the perimeter of economically viable uses in a context where demand is neither saturated nor constrained by explicit energy budgets. Usage volumes have grown by an estimated factor of 50 to 100 between 2022 and 2025, far outweighing unit efficiency gains.

This dynamic recalls Wirth's law in computing: "software slows down faster than hardware accelerates." In the AI context, one can propose an equivalent heuristic: models consume faster than chips economise. This is not a physical law, but a regularity observed in a regime where efficiency unlocks new degrees of freedom rather than reducing total consumption.

5.2. Agentic AI as catalyst of the rebound effect

The emergence of so-called agentic systems, autonomous agents capable of orchestrating complex tasks, maintaining extended context, and operating continuously, acts as a catalyst for this dynamic. The energy difference between paradigms is structural. A one-off query mobilises resources for a few

seconds. An autonomous agent operates over long horizons: persistent in-memory contexts, regular queries to external systems, coordination of several specialised models, continuous feedback loops.

At constant installed GPU infrastructure, the shift to persistent agents raises the effective utilisation rate (duty cycle) of resources from 15–30 percent (ad-hoc human usage) to 60–80 percent (24/7 agents). This intensification translates into an absolute energy consumption increase of ×3 to ×5, not because the models become less efficient, but because they are mobilised permanently.

The paradox is temporal. In the short term, agentic AI helps resolve an economic problem: the monetisation of already-financed infrastructure. In the medium term, it accelerates the collision with irreducible physical constraints. AI encounters its limits not in spite of its efficiency gains, but because of the way those gains expand the field of possible uses.

VI. Recomposition of competitive advantage

6.1. Evolution of success factors

In the early 2020s, competitive advantage in AI rested mainly on intangible assets: scientific talent, architectural innovation, and access to vast proprietary datasets. Five years on, these factors remain necessary, but their relative weight has been redistributed. The rapid diffusion of knowledge through open source, preprints, and talent mobility reduces the durability of purely algorithmic advantages. Foundation models are progressively converging in performance across a broad spectrum of tasks.

In parallel, factors hitherto secondary are gaining importance: access to abundant low-carbon energy, secured water resources, proximity to electricity generation infrastructure, territorial acceptability of projects, and, increasingly, access to high-quality proprietary data corpora in vertical domains where public data is insufficient. These dimensions belong more to the logic of heavy industry than to that of software.

6.2. AI as heavy industry

This redistribution brings AI structurally closer to industries historically constrained by access to energy and natural resources. Aluminium concentrated around hydroelectricity at the beginning of the twentieth century, heavy chemistry near ports, water, and electricity, and semiconductors have long been water-intensive and geographically concentrated.

The energy intensity of AI remains difficult to measure precisely, due to the heterogeneity of architectures and uses. Nonetheless, order-of-magnitude comparisons suggest an intensity of 0.40 to 0.60 kWh per dollar of added value for generative AI (methodological note: this estimate includes inference, training, and cooling relative to revenues directly attributable to AI; it excludes hardware amortisation and datacenter construction costs), against roughly 0.05 for financial services, 0.08 for traditional software, 0.80 for steel, and 1.20 for aluminium. AI has not yet joined electrometallurgy, but it is moving structurally closer.

6.3. New determinants of competitive advantage

In this new regime, the sources of durable competitive advantage are being recomposed: long-term energy contracts (low-carbon PPAs over 10 to 20 years treated as a strategic asset); proximity to production sources (dispatchable and abundant energy); secured water rights (an advantage difficult to replicate in water-stressed regions); political and social acceptability; and control of the energy chain through partial vertical integration (advanced nuclear, SMRs, long-term partnerships).

VII. Geopolitics of AI under physical constraints

7.1. Fragmentation of advanced-compute value chains

The production of advanced compute processors rests on a highly fragmented and asymmetrically distributed value chain. Design remains predominantly Western, while the fabrication of advanced logic chips is quasi-monopolistic: Taiwan concentrates more than 90 percent of world capacity below 7 nm. HBM memory depends on a few actors concentrated in East Asia. EUV lithography remains a European industrial monopoly (ASML).

This architecture creates major systemic vulnerabilities. Taiwan constitutes a single point of failure: any durable disruption would block the bulk of world production of advanced chips. Beyond considerations of international law and democratic bastion, a Chinese military intervention in Taiwan would heavily disrupt this new AI-linked economy.

7.2. The China/West asymmetry under physical constraints

Since 2022, China has undertaken a massive effort of technological self-sufficiency, sometimes described as the “silicon Manhattan Project.” Real progress has been observed (7 nm etching by DUV workaround at SMIC), but at the cost of low yields, energy overheads, and uncertain scalability.

The implicit assumption that physical constraints apply symmetrically to all actors, however, deserves scrutiny. China enjoys significant structural advantages along several of the dimensions identified in this article. In terms of energy-infrastructure deployment, Chinese execution speed has no equivalent in the West: between 2020 and 2024, China brought online more solar capacity than the rest of the world combined, while maintaining a nuclear programme (150 reactors planned or under construction) that far exceeds combined European and American ambitions. In terms of social and territorial acceptability, the Chinese governance model eliminates the delays linked to local opposition, environmental permits, and public consultations that extend projects in Europe and North America by several years.

The Chinese strategy is not limited to lithographic workarounds. It includes massive investments in alternative architectures (chiplets, wafer-level integration, compute-in-memory) that could partially obsolete the nanometre race. Two scenarios emerge: either a gradual convergence toward advanced nodes, neutralising the Western technological weapon, or a bypass through different architectural paths. In both cases, China’s relative advantage on physical constraints (rapid energy deployment,

execution capacity, political tolerance of externalities) could partly compensate for its lag on advanced silicon.

7.3. HALEU uranium and new nuclear diplomacy

The rise of SMRs and advanced reactors, considered to power AI infrastructures, introduces a critical dependence on HALEU uranium (5 to 19.75 percent). The current market is extremely concentrated: Russia supplies the bulk of world production. In the context of the war in Ukraine, depending on Russia for HALEU supply would amount to financing the conflict by swapping oil for uranium. HALEU also poses an international governance problem: its enrichment level is close to the military threshold, drastically shortening the breakout time toward military use in the event of diversion.

7.4. Submarine cables and digital sovereignty

The materiality of the internet rests almost exclusively on submarine cables. Over the past decade, their ownership has shifted from telecom operators to hyperscalers. Routes are concentrated around a few geographical chokepoints (North Atlantic, Eastern Mediterranean, Asian straits), exposed to sabotage and geopolitical pressure. For AI, any durable fragmentation of networks would lead to a forced regionalisation of models, accentuating asymmetries between blocs.

VIII. Counter-forces and zones of uncertainty

Energy and water constraints do not constitute a strictly impassable horizon. Several dynamics are likely to bend the current trajectory. Their scope, however, remains uncertain, which calls for nuancing any deterministic reading, without invalidating the structural trends identified.

8.1. Efficiency innovations: real gains, limited systemic reach

Progress in algorithmic and hardware efficiency is real. Sparsification, quantisation (FP32 to INT8/INT4), Mixture-of-Experts architectures, and model distillation significantly reduce inference costs and memory consumption, sometimes by an order of magnitude. On the hardware side, promising trajectories are emerging (neuromorphic, photonic, in-memory computing). Hooker (2021) shows that the computational load required to reach a given performance on ImageNet was divided by 10 between 2012 and 2022. Nevertheless, these improvements have been accompanied by rapid growth in total usage volumes, confirming the dominance of the rebound effect. Moreover, most of these innovations involve long industrial deployment lead times (5 to 10+ years).

8.2. Decentralised inference and edge computing

An alternative trajectory is emerging, so far poorly integrated into energy-consumption analyses: decentralised inference on distilled or quantised models, executed on local devices. Apple Intelligence, on-device Llama models, and the deployment of NPUs (Neural Processing Units) in consumer processors (Qualcomm Snapdragon X, Apple M4, Intel Meteor Lake) sketch a scenario in which a significant fraction of inference migrates to the edge, reducing pressure on centralised datacenters.

This scenario does not solve the problem of training (which remains centralised and hyper-intensive), nor that of persistent agents (which require extended memory contexts and multi-model coordination incompatible with current NPU capabilities). It could, however, significantly modify the energy profile of interactive inference, which represents the majority of current operational consumption, by redistributing the load across billions of devices whose unit energy cost is negligible. The magnitude of this transfer will depend on the speed of convergence between the capabilities of on-device models and user expectations regarding response quality.

8.3. Compute scheduling and intelligent energy orchestration

A technological counter-force already in deployment deserves attention: intelligent compute scheduling, which aligns compute workloads with the temporal availability of decarbonised energy. Google pioneered this approach with its carbon-aware computing system, which shifts training and batch-processing workloads toward the hours and regions where available electricity is the least carbon-intensive.

This approach reduces the marginal carbon footprint without reducing absolute energy consumption. It is ineffective for low-latency workloads (real-time inference) but relevant for training, fine-tuning, and batch tasks, which represent a substantial fraction of total consumption. Its effectiveness, however, depends on the local energy mix: in predominantly fossil systems, shifting a workload from night to day (to capture solar output) may have only a marginal impact if baseload production remains carbon-intensive.

8.4. Dedicated renewables: conditional effectiveness

Hyperscalers invest massively in renewable energy. Microsoft has contracted more than 10 GW of capacity by 2030; Google targets 24/7 decarbonised energy supply. Their actual effectiveness, however, depends closely on the local energy mix. In electrical systems still predominantly fossil (as with several U.S. regional grids whose mix oscillates between 60 and 80 percent fossil fuels), the purchase of so-called “green” electricity does not necessarily correspond to additional physical consumption of decarbonised energy. Power Purchase Agreements and certificates of origin largely operate an accounting reallocation of existing production.

In the absence of a simultaneous addition of dispatchable low-carbon capacity (nuclear, hydroelectricity, geothermal) or long-duration storage solutions, the increase in datacenter-related demand translates mechanically into higher fossil production elsewhere in the system. “Greening” strategies may thus reduce the declared footprint without reducing, or even while increasing, global emissions at grid scale.

8.5. Emerging regulation and European positioning

Regulatory frameworks are evolving rapidly, but in a fragmented way. The European Union is debating a carbon tax on compute; California is considering transparency obligations on the water footprint;

Singapore imposes strict PUE norms; Ireland caps the share of datacenters in national electricity consumption.

European positioning deserves particular attention. The Gaia-X initiative, the French sovereign cloud projects (S3NS, Bleu), and the reflections around a European "cloud of trust" represent an institutional attempt to answer the asymmetries identified in this article. Their effectiveness, however, remains uncertain. The fragmentation of national initiatives, implementation delays, and the absence of advanced etching capacity in Europe all limit the reach of these strategies. Europe, on the other hand, holds real assets on physical constraints: one of the least carbon-intensive electricity mixes in the world (France, Sweden, Finland), advanced regulatory experience (AI Act, GDPR), and industrial know-how in nuclear. The question is whether these structural advantages will be mobilised at the scale and speed required.

The risk of carbon leakage is real: some jurisdictions may maintain permissive regulations to attract investment. No counter-force taken in isolation constitutes a unique solution. Their combination could push back the deadline of physical constraints, without removing them.

IX. Strategic implications

9.1. For technology companies

In the short term (2025–2026), the strategic priority is to secure energy access before a further tightening of markets. Long-term energy contracts become strategic assets on a par with patents or talent. Multi-site geographic diversification across complementary jurisdictions strengthens resilience. In the medium term (2027–2030), partial vertical integration into the energy chain (fusion, SMRs, TerraPower) marks a cultural rupture for a sector historically centred on software. Finally, cost/latency/footprint trade-offs will need to be explicitly formalised: workloads tolerating high latency shifted toward energy-advantaged sites, with real-time inference remaining close to end users.

9.2. For regulators

Regulators have a limited window of action, roughly between 2025 and 2027, before investments produce durable lock-in effects. Available levers include binding efficiency norms (PUE caps, obligations to reuse waste heat), sector-specific carbon markets applied to compute, territorial planning (zoning, regional quotas), and mandatory transparency (standardised and audited consumption reporting). Their effectiveness depends on international coordination: without minimal harmonisation, national policies risk neutralising one another.

9.3. For territories

Territories with comparative advantages face a strategic choice: to valorise these assets to attract massive investment, or to limit their exposure to an energy-intensive and potentially volatile industry.

The opportunities are real (billions in investment, skilled jobs, tax revenue). The risks are substantial (mono-industry dependence, stranded assets, use-conflicts).

The controversies surrounding mega-reservoirs in French agriculture offer an illuminating precedent. Although they rest on a rational hydrological logic, these infrastructures have crystallised intense social opposition, not on their technical feasibility, but on the perception of a private appropriation of a common resource. AI datacenters fit into a comparable dynamic. Without explicit mechanisms for local counterparties, governance, and environmental compensation, social acceptability becomes a dominant limiting factor, independent of the technical relevance of the project.

X. Conclusion: the thermodynamic question (and therefore political)

Artificial intelligence has crossed a threshold: what was presented as an "augmented software" industry now behaves like an electro- and hydro-intensive industry, territorialised, constrained by physical resources and therefore politically contested.

This regime change exposes a central contradiction. On one side, dominant actors promote AI democratisation in order to broaden the user base, accelerate adoption, and amortise the massive investments already committed. On the other, the materiality of constraints reintroduces the management of opportunity cost. When electricity, water, data, and network capacity become binding factors with rising marginal cost, not all uses are of equal worth. Mobilising a megawatt for scientific research, for health, for climate modelling, or for critical infrastructure does not have the same social return as mobilising the same megawatt for recreational or weakly productive uses.

The tension then becomes systemic. The logic of maximal democratisation, required for industrial ROI, enters into conflict with a logic of efficient allocation of scarce resources. In other words, the question is no longer only who can access the models, but *which uses* justify, in view of their collective value, the real marginal cost of the resources they mobilise.

The ultimate question is therefore no longer algorithmic alone. It is thermodynamic, and immediately political: what share of electricity, water, data, network, and social acceptability will our societies agree to allocate to AI, and at the expense of which other uses? The answer will be played out in infrastructure trade-offs, in the governance of externalities, and in the collective hierarchisation of legitimate uses of compute. AI does not escape the laws of physics: it forces us to decide, explicitly, how we choose to pay them.

References

Arthur, W. B. (1989). Competing technologies, increasing returns, and lock-in by historical events. The Economic Journal, 99(394), 116–131.

Jevons, W. S. (1865). The Coal Question: An Inquiry Concerning the Progress of the Nation, and the Probable Exhaustion of Our Coal Mines. Macmillan.

Hooker, S. (2021). The hardware lottery. Communications of the ACM, 64(12), 58–65.

Villalobos, P., et al. (2022). Will we run out of data? An analysis of the limits of scaling datasets in machine learning. arXiv:2211.04325.

Muennighoff, N., et al. (2023). Scaling data-constrained language models. Advances in Neural Information Processing Systems (NeurIPS).

Shumailov, I., et al. (2023). The curse of recursion: Training on generated data makes models forget. arXiv:2305.17493.

International Energy Agency (2024). Electricity 2024: Analysis and Forecast to 2026. IEA Publications.

World Resources Institute (2023). Aqueduct Water Risk Atlas. WRI Global.

Boston Consulting Group (2024). From Potential to Profit with GenAI: Bridging the Gap. BCG Henderson Institute.

McKinsey Global Institute (2025). The State of AI: How Organizations Are Rewiring to Capture Value. McKinsey & Company.

Gallup (2024). The Real State of AI Adoption in American Workplaces. Gallup Inc.

US Census Bureau (2025). Business Trends and Outlook Survey: Artificial Intelligence Use. BTOS Q2 2025.

Stanford HAI (2024). Artificial Intelligence Index Report 2024. Stanford University.

European Parliament (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act).

Sorrell, S. (2009). Jevons' Paradox revisited: The evidence for backfire from improved energy efficiency. Energy Policy, 37(4), 1456–1469.

Kaack, L. H., et al. (2022). Aligning artificial intelligence with climate change mitigation. Nature Climate Change, 12, 518–527.