OCTOBER 2025

POLICY PAPER

THE ENERGY ECONOMICS OF ARTIFICIAL INTELLIGENCE IN A FRACTURED GLOBAL SYSTEM

RIM BERAHAB

POLICY CENTER FOR THE NEW SOUTH

PP - 37/25

POLICY CENTER FOR THE NEW SOUTH

Artificial intelligence (AI) is rapidly emerging as both an energy optimizer and a structural source of energy demand. While AI promises efficiency gains in forecasting, grid management, and emissions reduction, its expansion is already reshaping electricity systems: data center consumption could more than double by 2030. Beyond this techno-economic duality lies a deeper challenge: the sovereignty of digital and energy systems. AI rests on highly concentrated supply chains of chips, compute infrastructure, and critical minerals, as well as on access to abundant, low-carbon electricity. This concentration creates new dependencies and asymmetries, reinforcing the strategic control of a handful of actors. For Africa, the stakes are particularly high. The continent holds significant reserves of cobalt, manganese, rare earths, and other inputs indispensable to batteries and semiconductors, yet faces chronic electricity deficits, fragile grids, and limited compute capacity. Without deliberate investment in infrastructure, regional integration, and industrial upgrading, Africa risks remaining confined to raw-material supply while depending on foreign actors for digital infrastructure and cloud services.

RIM BERAHAB

** Many thanks to Ilias Dekkoun, expert in digital transformation and AI, for his insightful comments and suggestions.*

INTRODUCTION

Artificial intelligence (AI) is rapidly evolving into a general-purpose technology—much like mechanization, electricity, or the internet—in terms of its potential to lead to a reconfiguration of production systems, economic organization, and geopolitical hierarchies. AI has a long history dating back to the 1950s. Over time, it has been through alternating periods of optimism and pessimism—the latter dubbed 'AI winters'. Recently, its development has been enabled by four main drivers: exponential improvements in computation power and cost, a surge in digital data availability, an increase in internet speed, and a series of algorithmic breakthroughs, particularly in machine learning and neural networks.

What began as a frontier of academic experimentation seems now to become an industrial juggernaut: the combined market capitalization of AI-related firms in the S&P 500 has increased by more than $12 trillion since 2022, drawing in huge capital, leading to recalibration of corporate strategies, and reshaping national policy priorities (IEA, 2025). While some valuations may reflect hype rather than immediate profitability (Observer, 2025), AI focused start-ups had an average valuation five times higher than other start-ups, triggering in turn a surge in investment in AI-related infrastructure. In particular, Alphabet, Amazon, Meta, and Microsoft planned as much as $300 billion in AI-related capital expenditure in 2025, 20% higher than total power sector investment in the United States (IEA, 2025).

Amidst the hype around AI, a fundamental truth remains: AI runs on energy. Every data point processed, every algorithm trained, and every inference made requires power, often in vast, concentrated quantities. This huge demand is serviced by energy-intensive data centers with electricity consumption that rivals that of mid-sized economies. The cooling systems that keep servers operable, the semiconductors that execute trillions of operations per second, and the global infrastructure linking these systems all depend on affordable and reliable energy, which is projected to become increasingly clean. In short: there is no AI without energy.

AI is also being heralded as a transformative force within the energy sector itself. It promises to optimize grid operations, forecast renewable energy availability, enable dynamic demand management, and even guide infrastructure planning. In theory, this creates a virtuous cycle: AI needs energy, and AI can help manage and decarbonize energy systems. The question, however, is whether the efficiency gains AI enables can offset the surge in energy consumption it generates. Striking this balance—between AI's economic promise and its mounting energy footprint—will be critical to determining whether AI becomes an accelerator of the energy transition, or an additional source of strain on already pressured energy systems. The deepening interdependence between AI and energy remains poorly understood and underexplored in policy and academic circles, especially in emerging economies in Africa.

Yet the rise of AI is not only a techno-economic challenge. It is also a question of sovereignty. AI depends on highly concentrated supply chains of compute hardware, cloud services, and critical minerals, alongside abundant, low-carbon electricity. This concentration creates new dependencies and asymmetries, reinforcing the strategic control of a handful of actors. For late-industrializing regions such as Africa, where electricity access remains limited but mineral endowments are significant, the AI–energy nexus crystallizes both risks of exclusion and opportunities for strategic positioning.

This paper tackles a central issue: how does the rise of AI, as both an energy-consuming force and an energy-optimizing tool, reshape the economic, industrial, and policy logics of the global energy sector, and what does this mean for late-industrializing regions such as Africa?

I. UNDERSTANDING ARTIFICIAL INTELLIGENCE (AI)

AI is increasingly described as the defining technological paradigm of the twenty-first century. AI is not just a software development; it is a computationally and materially intensive system that rests on a vast infrastructure of energy, hardware, and data. This section provides a conceptual framework to understand what AI is, and how it works.

AI refers to the development of systems capable of performing tasks that typically require human intelligence: predicting outcomes, recognizing patterns, interpreting language, or making decisions. From a technological and engineering perspective, the development of AI has involved three phases:

1. **Rule-based or symbolic AI.** This was the earliest form of AI, in which machines follow explicitly programmed instructions and logical rules. A chess engine programmed to evaluate every legal move using a decision tree falls into this category. These systems are rigid and hard to scale because all scenarios must be anticipated and coded by humans.
2. **Machine learning (ML).** In contrast to rule-based systems, ML models learn from large datasets. Instead of being told exactly how to perform a task, these algorithms identify patterns and correlations in the data to generate predictions. For instance, a fraud-detection model can learn to flag unusual credit-card activity by analyzing millions of historical transactions.
3. **Deep learning (DL).** This is a subset of machine learning that relies on artificial neural networks. Inspired by the human brain, these models consist of multiple layers of interconnected 'nodes' that process and transform inputs. Deep learning models are capable of handling unstructured data such as images, audio, and natural language, and are behind recent advances in AI applications, such as ChatGPT and facial-recognition systems.

In addition to this historical-technological lens, AI can also be classified based on the functional nature of the tasks it performs. This helps policy analysts and energy planners assess AI's systemic impact not just by how models are trained, but also by where they operate — whether in the physical world (e.g., robots, grid hardware), the digital space (e.g., forecasting models, data analysis), or in hybrid systems that combine both (e.g., digital twins of energy assets) (IEA, 2025):

- Predictive AI refers to systems designed to forecast future outcomes by identifying patterns in historical and real-time data. Its computational backbone is typically composed of statistical models, time-series forecasting, and machine learning algorithms. Its applications encompass energy-demand forecasting, weather prediction, predictive maintenance of energy infrastructure, and financial modeling.
- Generative AI systems, which are pretrained on vast amounts of data to recognize and replicate patterns, are capable of producing content—text, images, video, or code. Generative AI can be further decomposed into: (i) language models such as GPT, which process and generate text, (ii) multimodal models capable of handling and producing content in multiple formats, such as text-to-video generation, and (iii) Large Reasoning Models, which are advanced large language

models (LLMs), optimized for step-by-step reasoning (such as OpenAI's o1 or DeepSeek's R1).

- Computer vision, which enables machines to interpret and analyze visual data, such as images and video feeds. Its applications include facial recognition, object detection, industrial surveillance, autonomous navigation, and optical character recognition, which converts printed or handwritten text into machine-readable formats (like Google Vision).
- Physical AI, or embodied AI, refers to systems that physically interact with their environment, often combining robotics and AI-based perception and control. Its applications include autonomous vehicles, robotic arms, and drones.
- Agentic AI refers to AI agents that perform autonomous tasks in digital or physical environments, often interacting with humans or other agents. Its mains use include workflow automation, customer support bots, and smart building energy and management systems.

These functional categories are not mutually exclusive. A single AI application—such as an autonomous electric vehicle—can integrate generative (language interface), predictive (navigation), and physical AI (vehicle control).

Governments view AI as a general-purpose technology, much like electricity or the internet, that can transform multiple sectors simultaneously. However, unlike many digital tools, AI systems require a significant amount of physical infrastructure and, by extension, energy. The functionality of AI rests not only on software algorithms but on energy-intensive computation and hardware. This raises sovereignty considerations: countries can either rely on AI delivered as a service through global cloud providers—accepting a degree of dependence—or invest in developing the capacity to run AI in-house, with greater control over data, energy use, and strategic autonomy, but at the cost of heavy infrastructure investments, high operating expenses, and the risk of rapid technological obsolescence. Understanding this trade-off is critical for policymakers who seek to regulate, adopt, or strategically invest in AI systems.

II. AI IN THE ENERGY SECTOR: OPPORTUNITIES AND CONSTRAINTS

AI is emerging as a pivotal tool in transforming the global energy system. The sector's complexity, stemming from interdependent flows between primary supply, transformation, and end-use, offers fertile ground for AI-driven optimization. This complexity is deepening because of structural trends in the energy sector, including electrification, digitalization, decentralization, and the rapid integration of variable renewables. AI, by design, is well suited for managing such complexity.

When harnessed effectively, AI applications in the energy sector can be grouped into two types: first, facilitating development by identifying resources and designing, planning, and building facilities; second, optimizing, refining, and automating the operation of energy systems (IEA, 2025). However, the current state of AI deployment in energy is still at an early stage and uneven. This section assesses how AI is being, or could be, applied—focusing on two applications in the oil and gas industry and electricity systems—and identifies the barriers limiting its broader uptake.

2.1. Oil and Gas: Enhancing Safety, Reducing Emissions, and Lowering Costs

The oil and gas industry has long served as a proving ground for advanced technologies, and AI is no exception. As exploration and production costs rise due to geological and environmental challenges, firms have adopted increasingly sophisticated computational tools. Supercomputing capacity within the oil and gas sector has nearly doubled since 2010, growing at approximately 70% annually (IEA, 2025). AI tools are being deployed across both upstream and midstream segments, to boost efficiency and reduce costs.

Exploration and Development

AI could play an increasingly important role in exploration and development. Subsurface exploration has traditionally relied on seismic surveys and reservoir modelling, both of which generate extremely large and complex datasets. Interpreting these data with conventional methods takes time and is prone to uncertainty, often requiring multiple exploratory wells before confirming the presence of commercially viable reserves. AI tools, by contrast, can detect patterns in seismic echoes and geological formations that human analysts or traditional algorithms might miss. In practice, this means that AI-enhanced seismic processing can classify underground structures with much greater reliability. Studies suggest up to 90% greater accuracy from AI tools in distinguishing oil- and gas-bearing formations from non-productive rock (Araya-Polo et al, 2017). For decision-makers, the implication is straightforward: fewer unnecessary wells are drilled, exploration costs fall, and development risks decline.

Beyond initial exploration, AI is transforming reservoir simulation. Traditionally, modeling the physical properties of rocks and fluids underground—such as how oil, gas, and water move under pressure—required months of computation. Advances in physics-informed machine learning now allow these models to integrate seismic data, well logs, and production histories in near real time. A process that once took months can now be completed in hours, with a greater degree of precision. Chevron has already applied such AI-driven simulations to improve well placement and optimize production forecasts (JPT, 2022). It recorded a 30% increase in drilling speed, with a corresponding 25% reduction in operational costs, resulting from AI-driven automated drilling (Adjei et al., 2025).

Operation and Safety

Once production is underway, AI can support a shift toward more digitalized and automated operations. Oilfields today are equipped with tens of thousands of sensors generating terabytes of data on pressure, temperature, flow rates, and equipment performance. AI enables real-time analysis of these streams, allowing operators to anticipate problems before they occur through predictive maintenance, and helping them monitor facilities remotely, an especially valuable capability in deepwater or otherwise hazardous environments.

Among the most significant advances is the development of digital twins, which are virtual replicas of physical assets, such as wells, pipelines, or entire offshore platforms, which are continuously updated with live operational data. These digital twins allow operators to stress-test systems in a virtual environment, simulate failures, and optimize performance without interrupting actual operations. By minimizing unplanned downtime and enabling remote supervision, they can reduce costs while lowering safety risks by reducing the need for on-site intervention.

AI-driven forecasting tools have also advanced significantly. Hybrid models, which combine machine learning with traditional engineering approaches, now outperform purely statistical methods, reducing forecasting errors by as much as 25% in certain applications (Kuang et al, 2021). For deepwater projects—for which risks are high and margins tight—such efficiency gains can be significant, though precise financial figures are rarely disclosed due to commercial sensitivity, suggesting that such claims should be treated with caution.

Emissions Reduction and Carbon Management

One of the most strategically significant frontiers for AI in the energy sector lies in its ability to transform emissions control. Fossil-fuel operations remain a major source of global greenhouse gases, with methane alone accounting for roughly one-third of anthropogenic emissions. Methane is far more potent than CO_2 in the short term, but its emissions are often invisible and dispersed, making them difficult and costly to track using conventional methods.

AI has the potential to change this equation. By combining advanced algorithms with satellite imagery from platforms such as Sentinel-2, operated by the European Space Agency (ESA) in Europe, and Landsat, operated by the United States Geological Survey (USGS) in the United States[1], companies can monitor oil and gas operations almost continuously. This allows for the rapid detection of methane leaks across wide areas, reducing both the time and the expense required for inspections. For policymakers, this means that leak detection can shift from being a reactive, labor-intensive process, toward a proactive, real-time monitoring system. The climate dividends are potentially substantial: faster detection translates directly into quicker repairs, curbing one of the most damaging sources of emissions (Xia et al, 2024).

The promise of AI extends beyond identifying leaks. Deep learning models can be trained on historical operational data to predict where leaks are most likely to occur—whether because of aging equipment, fluctuating pressure levels, or weak seals. This proactive capacity enables operators to respond more effectively, and also to design maintenance schedules and safety protocols that minimize risks before they materialize. In this sense, AI supports a structural shift from 'detection and repair' toward 'anticipation and prevention', which is critical for aligning fossil-fuel operations with global climate goals.

In parallel, AI is beginning to shape carbon capture, utilization, and storage (CCUS) systems, which are central to long-term decarbonization strategies. Effective CCUS depends on precise modeling of underground reservoirs to ensure that injected CO_2 remains securely stored. Although the sector is still nascent, several pilot and early commercial projects are underway. For example, TotalEnergies plans to capture nearly 1 million tonnes of CO_2 annually at its Papua LNG project, with operations expected to begin in late 2027, underscoring the technology's emerging status (TotalEnergies, 2022).

Despite the limited number of operational CCUS sites relative to global fossil fuel infrastructure, AI is already being used to enhance predictive modeling, monitor storage integrity, and guide investment decisions. The partnership between TotalEnergies and AI hardware firm Cerebras illustrates this trend. The Cerebras CS-2 system delivers more than 200× faster performance for

[1] These satellites cover the entire globe and provide open-access data. Their capabilities have enabled near-continuous monitoring of oil and gas operations worldwide by detecting atmospheric pollutants, methane emissions, and other environmental indicators.

complex reservoir simulations compared with traditional GPUs, improving confidence in storage security, optimizing deployment strategies, and potentially lowering costs (Cerebras, 2022). While this collaboration highlights tangible advances in applying cutting-edge AI to CCUS, it remains at an advanced R&D stage rather than widespread operational use. AI's role in CCUS is therefore promising but still at an emerging stage.

2.2. Power Systems: From Forecasting to Grid Optimization

If in the oil and gas sector AI is primarily about lowering production costs and tightening emissions control, its role in power systems is far more systemic: it is about making the entire electricity grid fit for an era of high renewable penetration and accelerating electrification. Unlike hydrocarbons, for which the challenge lies in extracting and managing finite resources, power systems must handle unprecedented levels of variability and complexity. The shift from centralized fossil-based generation to distributed solar, wind, and storage has upended traditional grid architectures and exposed new vulnerabilities, but has also created opportunities for optimization. In parallel, the volume of system-relevant data—on demand patterns, weather conditions, equipment performance, and market signals—has exploded, creating both a challenge and an opportunity. AI's capacity to analyze, predict, and optimize in real time positions it as a central enabler of next-generation electricity systems.

Forecasting, Balancing, and System Flexibility

Accurate forecasting is the foundation of a renewable-heavy power system. AI-enhanced models can integrate real-time meteorological data, historical consumption, and localized grid information to deliver far more precise short- and medium-term forecasts. This capability reduces the need for costly reserve margins, and minimizes renewable curtailment. For example, Google DeepMind's wind power tool increased the value of wind generation by 20% by providing more accurate short-term forecasts of output (Google DeepMind, 2019). Similarly, National Grid ESO in the United Kingdom uses AI to improve solar forecasts up to eight hours ahead (Fulton et al., 2024), while RTE in France and Elia in Belgium deploy AI in real-time operations to anticipate system imbalances (IEA, 2025).

According to the IEA, even a small reduction in the amount of renewable electricity that goes unused—known as curtailment—can have a major global impact. Cutting average curtailment by just one percentage point in 2035 would avoid fossil-fuel demand equivalent to 28 million metric tons of coal and 14 billion cubic meters of natural gas, preventing around 120 Mt of CO_2 emissions (IEA, 2025). This shows that AI-driven forecasting and grid flexibility can also be critical tools for scaling renewables, reducing waste, and strengthening energy security.

Planning, Permitting, and Construction Efficiency

Beyond operations, AI can also reshape the way renewable projects are planned and built. In the design phase, AI models can simulate site conditions and optimize layouts, reducing both land use and transmission costs. Tools such as the U.S. National Renewable Energy Laboratory's Wind Plant Graph Neural Network or Iberdrola's Sedar project illustrate how AI can fine-tune turbine spacing and solar orientation to maximize performance.

Permitting, often a major bottleneck in renewable deployment, is another area of innovation. In the United States, AI systems trained on over 28,000 environmental impact documents now help accelerate approval processes, cutting review times that can stretch into years (PNNL, 2024). During construction, AI-based logistics optimization, such as GE Vernova's models targeting a 10% reduction in turbine installation costs (GE Vernova, 2022), can help streamline supply chains, shorten build times, and lower capital expenditure.

Operations, Maintenance, and Asset Optimization

Once energy assets are operational, AI can enhance their reliability and extend their lifespan by analyzing sensor data in real time to detect anomalies and anticipate equipment failures before they occur, thereby reducing costly downtime. Companies are actively piloting such tools: Iberdrola's MeteoFlow and Enel's Myst AI, for instance, integrate weather-informed forecasts to optimize renewable output in real time (Iberdrola, 2016; Enel, 2022). Similarly, Hitachi Energy markets AI-driven image recognition tools to help transmission operators monitor equipment health and vegetation encroachment (Hitachi Energy, 2024), while in China, State Grid has reported using reinforcement-learning algorithms to optimize operational scheduling under complex trade-offs.

Independent assessments, however, suggest that many of these applications remain in early stages and their broader impact on cost savings is still being tested. A conservative IEA estimate indicates that AI deployment in operations and maintenance could reduce costs by up to 10%, equivalent to USD 40 billion annually (IEA, 2025). This points to substantial potential but also underscores the need for further evidence beyond company-reported gains.

2.3. Barriers to Systemic Integration: Infrastructure, Regulation, Trust

Despite its transformative potential, AI deployment across energy systems remains limited, uneven, and largely concentrated in advanced economies and isolated segments of the value chain. This gap stems not from a lack of technological potential, but from a confluence of structural, institutional, and sector-specific barriers that risk holding up systemic integration. Understanding these interlinked barriers is critical to designing a credible roadmap for systemic integration.

Infrastructure Deficits and Data Gaps

AI deployment rests on the availability of dense sensor networks, high-frequency data streams, and reliable cloud computing infrastructure. Yet, much of the world's energy capital stock was built decades before digitalization was conceivable. Oilfields, power plants, and transmission grids, particularly in emerging and developing economies, often lack the instrumentation and connectivity required for real-time monitoring. Retrofitting these assets is technically feasible but capital-intensive, and in many cases economically irrational given their long lifespans and the sunk costs.

Even where infrastructure exists, energy data is frequently fragmented, siloed in proprietary formats, or withheld due to commercial confidentiality or national security concerns. The absence of common standards undermines interoperability, limiting the scope for generalizing AI tools across systems. This curtails the potential efficiency gains in electricity grids or renewable energy integration, and slows the learning cycles necessary for AI systems to evolve and improve.

Institutional Inertia and Legacy Business Models

The oil and gas sector illustrates how technological ambition collides with institutional conservatism. Supercomputing and physics-informed machine learning have been deployed by a handful of majors to improve seismic imaging or reservoir modeling, yet most national oil companies and mid-sized firms remain constrained by long investment cycles and risk-averse cultures. Legacy assets, often capital-intensive and decades old, are rarely designed for digital retrofits. For utilities and grid operators, similar inertia is seen in operational models that are still geared toward centralized generation, even as renewables and distributed resources proliferate. Without structural incentives, institutions tend to preserve existing business models rather than experiment with digital innovation.

Regulatory Lag and Misaligned Incentives

Regulation has not kept pace with technological change. In oil and gas, unresolved questions around data ownership, liability for AI-driven failures, and the auditability of black-box systems discourage deployment. In the power sector, regulatory models continue to reward capacity expansion and cost-of-service, rather than efficiency gains or predictive capabilities that AI could unlock. Many regulators also lack the technical expertise to assess algorithmic forecasts or validate probabilistic models, creating a credibility gap that slows approval. For renewables, protracted permitting processes, opaque tariff-setting, and the absence of clear grid-access rules reinforce delays, making it difficult to implement AI-enabled solutions to accelerate planning or optimize siting.

Human Capital and Organizational Readiness

The shortage of professionals at the intersection of AI, engineering, and energy operations is another structural bottleneck. Advanced firms, particularly large oil majors or global technology providers, can attract scarce talent, but public utilities, regulators, and national oil companies often cannot. This asymmetry entrenches dependence on external vendors and narrows the scope for domestic capacity-building. Beyond skills, embedding AI requires organizational change and the development of a new culture: new procedures for data governance, procurement, and operational accountability. Many institutions remain unprepared to absorb these shifts.

Trust, Explainability, and Mission-Critical Systems

For mission-critical infrastructure such as electricity transmission or upstream oil operations, trust and accountability are paramount. Current AI systems, while powerful in pattern recognition, are often unable to explain their decision logic in ways that regulators and operators require. Transmission system operators, for instance, remain reluctant to entrust real-time grid balancing to AI tools, confining their use largely to planning or predictive maintenance. IEA (2025) surveys show that only 23% of transmission system operators currently deploy AI for real-time operations, a testament to the importance of explainability and human oversight in high-stakes environments. Taken together, these barriers form a self-reinforcing cycle: weak infrastructure creates fragmented data, which limits regulatory oversight. Regulatory gaps discourage investment, and institutional inertia prevents organizational change. In such a context, it cannot be assumed that AI will diffuse naturally. Its wide integration into energy systems requires deliberate policies: regulatory sandboxes to test new applications, performance-based incentives to reward efficiency, and public-private partnerships to de-risk early adoption.

III. THE ENERGY COST OF AI TODAY: A LIFECYCLE PERSPECTIVE

The previous section highlighted the potential of AI across energy systems, but also the barriers that hinder deployment. One crucial, often overlooked barrier is AI's own energy footprint. While AI can help optimize grids, decarbonize power systems, and improve efficiency, it is simultaneously adding a fast-growing layer of electricity demand. To understand this paradox, it is essential to examine AI not only as a tool for energy, but also as a consumer of energy, across its entire lifecycle. AI's energy consumption spans three key stages: hardware manufacturing, model training, and model usage (inference). These stages are sequential and cumulative: the substantial energy consumed in processing the raw materials and producing chips (upstream) is carried forward into the training phase (midstream), and the deployment of models at scale (downstream) transforms episodic consumption into a continuous demand driver. Taken together, they amount to a substantial and rising share of global electricity use.

Upstream: Hardware Manufacturing

The lifecycle begins with the extraction of critical raw materials and the fabrication of specialized chips, particularly graphics processing units (GPUs), which underpin modern AI. These are distinct from traditional central processing units (CPUs), which are general-purpose chips found in most computers. CPUs excel at sequential processing, suitable for routine tasks like web browsing or word processing. In contrast, GPUs, originally developed for rendering graphics in video games, are designed for parallel computation, enabling them to perform thousands of simultaneous calculations, which is essential for training complex AI models and handling large datasets efficiently.

Manufacturing these chips, particularly the advanced GPUs used for AI training, is one of the most energy-intensive industrial processes in the AI value chain. The production of a single 3-nanometer chip wafer, a leading-edge technology used in cutting-edge AI hardware, consumes approximately 2.3 megawatt hours (MWh) of electricity, according to Garcia Bardon et al (2021), roughly equivalent to the electricity an average U.S. household uses over two and a half months (EIA, 2022). This is due to the extreme precision and environmental control required during fabrication. Roughly 60% of the energy consumed during the production process goes into etching, layering, and photolithography processes, while the remainder is consumed for stabilizing the environment (cleanroom infrastructure, temperature stabilization, and support systems).

Therefore, the cumulative energy footprint of this upstream stage is far from marginal. Estimates by Greenpeace (2023) suggest that the semiconductor industry consumes over 100 terawatt hours (TWh) of electricity annually, roughly equivalent to the total electricity consumption of the Netherlands (110 TWh in 2024), or 1% of global industrial demand. As AI development continues to push toward smaller, denser chips and more compute-intensive architectures, this energy intensity is likely to increase.

Importantly, the energy cost is frontloaded, incurred well before an AI model is trained or deployed. The resulting chips embed a significant carbon and energy debt at the very start of the AI lifecycle. For policymakers and investors, this implies that any analysis of AI's environmental or energy impact must begin not at the data center, but in the chip foundry.

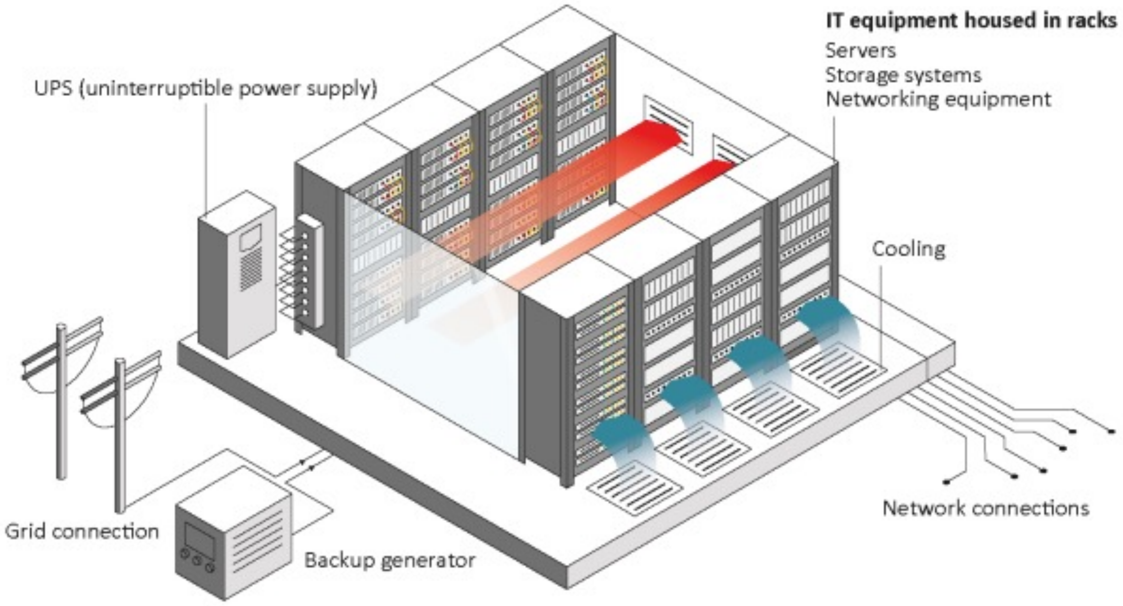
Midstream: Model Training

Once the hardware is manufactured, the next stage in the AI lifecycle is model training. This is the process through which an AI algorithm learns to identify patterns, relationships, and structures within massive datasets. The more parameters a model has and the more data it is trained on, the more computational resources, and therefore electricity, it requires.

Model training is not performed on ordinary devices. It takes place in data centers, which are centralized facilities that house racks of computer servers, networking systems, data storage units, and cooling infrastructure. These buildings provide the physical environment necessary for running high-performance computing operations continuously, securely, and at scale. Among them, hyperscale data centers are a specific category designed to handle vast volumes of computation and data. These facilities can contain hundreds of thousands of servers, connected through ultra-fast fiber-optic networks, and are often operated by global cloud service providers, such as Amazon Web Services, Microsoft Azure, and Google Cloud. They also require cooling systems, backup power systems and backup data servers (Figure 1).

Figure 1

Data Center Components



Source: IEA, 2025

For illustration, training GPT-4[2] reportedly involved using 25,000 GPUs operating at a combined rated power of 10 megawatts (MW). With a load factor of 84% and a training duration of 14 weeks (EpochAI, 2024), this corresponded to a total energy consumption of approximately 42.4 gigawatt hours (GWh), or about 0.43 GWh per day of training. That is equivalent to the daily electricity use of 28,500 households in advanced economies—or nearly 70,500 households in lower-income countries (IEA, 2025). This is not an isolated case. As of 2024, models like Gemini 1.5, Claude 3, and Meta's Llama 3 are expected to push compute and training requirements even higher, with

[2] This paper was written before the launch of GPT-5.

training datasets reaching the scale of trillions of tokens and model sizes in the hundreds of billions of parameters.

The carbon intensity of this process depends heavily on where the data center is located and what kind of electricity mix is used. A model trained on a coal-powered grid has a vastly different footprint to one trained in a hydropower-abundant region. Yet these differences are often invisible to end users. From an economic standpoint, this stage embeds significant operational costs, making access to cheap, stable electricity a competitive advantage. It also shifts part of the burden of AI expansion from innovation to infrastructure, placing strain on power systems and public utilities, especially in regions where grids are already near capacity.

Downstream: Model Inference and Use

Once an AI model is trained, it enters the phase of inference: applying what it has learned to real-world data in applications such as search queries, language translation, content generation, or autonomous decision-making. While the per-unit energy cost of inference is significantly lower than training, its cumulative energy impact can be larger, because of scale, frequency, and latency constraints.

The total inference load depends on:

- Model size: larger models require more computations per query.
- User base: mass deployment across billions of users, such as ChatGPT or Copilot, multiplies energy demand.
- Query length and complexity: a brief prompt may take milliseconds; generating video or performing logical reasoning may take seconds to minutes of GPU time.
- Hardware deployment: inference conducted on the edge (mobile phones, laptops) typically consumes less energy per task than inference performed on central cloud-based GPUs—but at the cost of performance.

According to Epoch.AI (2025), a typical ChatGPT query using GPT-4o likely consumes roughly 0.3 watt-hours (less than the amount of electricity that an LED lightbulb or a laptop consumes in a few minutes), which is ten times less than the estimate from 2023, suggesting positive development in terms of model efficiency and improved hardware. Nonetheless, inference is a continuous, recurrent process, unlike training, which is episodic. As models become embedded in critical systems—healthcare diagnostics, smart grids, financial trading, supply-chain optimization—the energy demand becomes persistent and systemic, raising new questions for grid planning, peak-load forecasting, and demand management.

IV. AI'S RISING ELECTRICITY DEMAND: SCENARIOS AND TRADE-OFFS

The previous section traced the lifecycle energy costs of AI, showing how each stage embeds significant electricity needs. Yet, while these costs are substantial in isolation, their true impact emerges when aggregated across global deployment. What was once a marginal load on power systems is rapidly evolving into a structural driver of electricity demand. AI is no longer only a tool for energy optimization; it is also becoming one of the fastest-growing sources of demand within the energy system, with implications for grid adequacy, investment planning, and net-zero strategies.

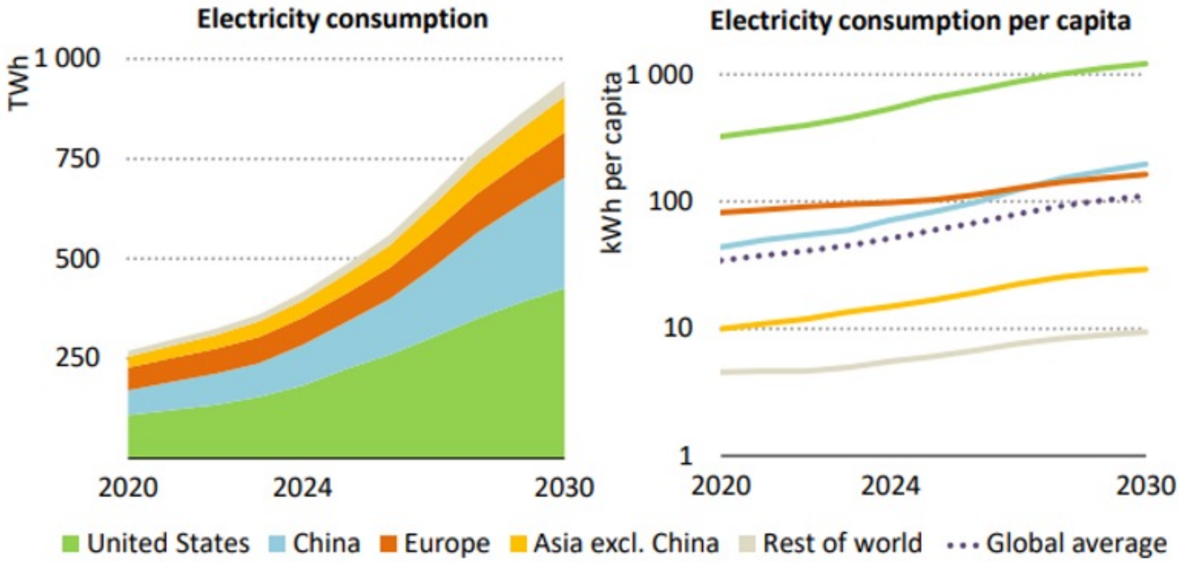
According to the IEA (2025), total electricity consumption of data centers, AI, and crypto combined could more than double from 460 TWh in 2022 to over 1,000 TWh by 2030. Most of this projected growth stems from the increasing energy footprint of AI workloads, both in training large models and running them (inference) at scale. In the United States, for instance, data center electricity consumption is projected to rise from 3% of total electricity use today to nearly 9% by 2030. Growth will be particularly acute in northern Virginia, already the world's largest data center hub due to its dense fiber networks, proximity to federal agencies, and concentration of hyperscalers, and in Texas, where abundant low-cost renewable energy and a deregulated electricity market attract large-scale data center deployments. Similar pressures are emerging in Ireland, Singapore, and parts of China, where power grid congestion has already led regulators to slow or restrict the approval of new hyperscale data centers (IEA, 2025).

To assess the range of possible energy futures, the IEA (2025) outlines four scenarios capturing divergent paths for AI and data center electricity consumption. Projections for the base scenario are provided in detail to 2030, based on current policies and market trends. Beyond 2030, the analysis extends all four scenarios—including the base scenario—into an exploratory horizon up to 2035, to illustrate how demand could evolve under contrasting assumptions.

The base scenario assumes continued growth of energy demand by data centers under current regulatory conditions and industry projections. Electricity use by data centers globally would rise from 250 TWh to 945 TWh by 2030, amounting to nearly 3% of projected global electricity demand (Figure 2). The data center sector would grow by 15% annually, far outpacing other sources of energy demand. Notably, accelerated servers—used in AI applications—would grow by 30% per year, contributing nearly half of the net increase in electricity demand. By contrast, conventional servers would grow at 9% annually. When extended to the exploratory horizon of 2035, electricity use in the base case is projected to reach about 1,140 TWh, underscoring a sustained upward trajectory absent stronger efficiency or policy interventions.

Figure 2

Data Center Electricity Consumption and Data Center Electricity Consumption Per Capita by Region in the Base Case, 2020-2030



Source: IEA, 2025.

In this scenario, the geographical disparity in terms of AI-related energy demand will widen by 2030 (Figure 4). The United States and China together are projected to account for nearly 80% of global growth in data center electricity demand between 2024 and 2030. In absolute terms, the United States is expected to add approximately 240 TWh, a 130% increase from 2024 levels. China follows closely with a projected increase of 175 TWh, reflecting a 170% growth. Europe contributes a more modest 45 TWh to the global increment (up 70%), while Japan's demand rises by 15 TWh (up 80%). This pattern confirms the strategic clustering of AI infrastructure in countries that combine robust digital economies with relatively low-cost, stable electricity supplies. Yet such concentration also raises grid-level challenges.

Southeast Asia is also emerging as a high growth node. Data center electricity demand in the region is projected to more than double by 2030, driven by the establishment of regional hyperscale hubs in Singapore and Johor province (Malaysia). This development is underpinned by aggressive digitalization policies, regional cloud service expansion, and competitive investment incentives.

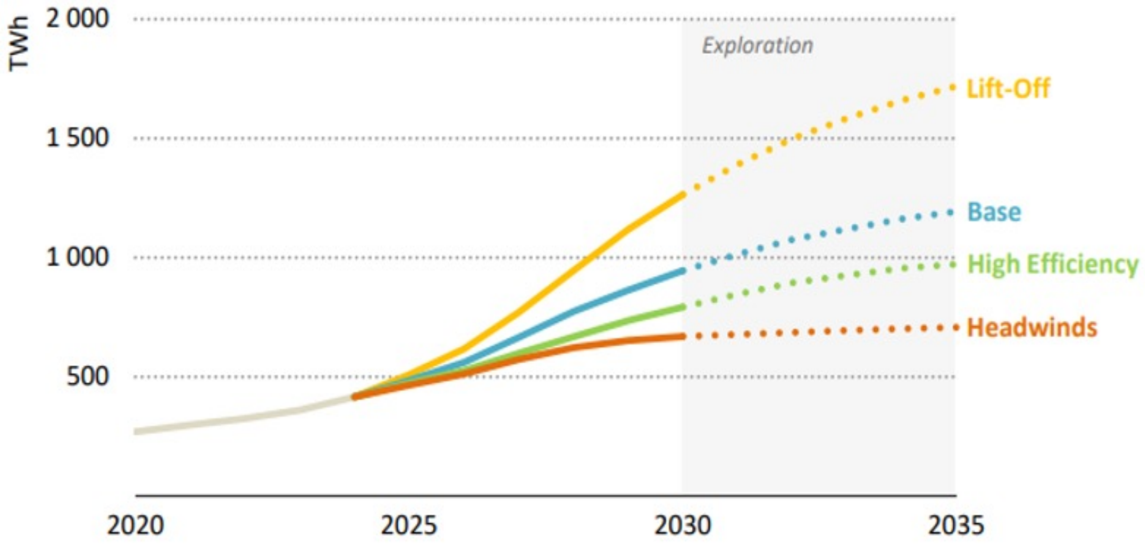
Looking at the landscape of data center electricity per capita provides a more revealing perspective on the weight of digital infrastructure in national energy systems. In 2024, the U.S. already led with about 540 kWh per person; this figure is projected to more than double to over 1,200 kWh by 2030. This intensity is an order of magnitude higher than any other region, highlighting the country's outsized AI-industrial footprint.

China's per-capita data center electricity use, at 70 kWh in 2024, is expected to triple to around 200 kWh by 2030, overtaking Europe, where the figure will rise more slowly from 100 kWh to 165 kWh over the same period. Japan will reach approximately 270 kWh per capita, consolidating its position as a high-density digital infrastructure economy. By contrast, India will remain at 15 kWh, underscoring its lower digital infrastructure energy intensity despite a large IT sector. Africa sits at the bottom of the distribution. In 2024, the continent's average per-capita data center electricity consumption was below 1 kWh, and is projected to rise only modestly to just under 2 kWh by 2030. However, this average masks significant variation: South Africa alone is expected to reach over 25 kWh per capita, more than 15 times the continental average, driven by localized cloud investment and industrial data demand.

The IEA compares the base scenario with three other scenarios (Figure 3). The 'lift-off' scenario assumes a stronger surge in AI adoption, driven by favorable financing and highly adaptable supply chains, notably increased chip availability. It also assumes rapid resolution of local constraints and increased reliance of on-site power generation, with grid infrastructure serving more as backup. In this scenario, global data center electricity use would reach 1,700 TWh by 2035, or about 4.4% of total global demand, a nearly 80% increase over the base case.

Figure 3

Global Data Center Electricity Consumption by Sensitivity Case, 2020-2035



Source: IEA, 2025.

The 'high efficiency' scenario assumes the same level of AI service demand as the base case but integrates ambitious energy-efficiency measures. These include code optimization, hardware improvements, and a migration away from smaller enterprise data centers toward colocation facilities and large-scale service provider infrastructure, particularly highly efficient hyperscale data centers. Compared to the base scenario, which projects global data center electricity consumption at around 1,140 TWh by 2035, the high efficiency pathway results in a 15% reduction, limiting demand to about 970 TWh (2.6% of global electricity use). This trajectory represents the most carbon-aligned outcome but depends on coordinated technical advances and supportive policy intervention.

Finally, the headwinds scenario assumes that AI adoption faces setbacks arising from monetization difficulties, public pushback, or supply chain delays. Growth in service demand is subdued, and power-sector constraints delay new installations. Despite improved energy efficiency, total demand flattens at 700 TWh, limiting data centers to under 2% of global electricity demand by 2035.

The scale and pace of AI-driven demand raise pressing questions about grid adequacy, infrastructure investment, and allocation of clean power. Unlike demand from electric vehicles, which is geographically dispersed, data centers are highly concentrated and require direct high-voltage access, stable generation capacity, and cooling resources. In the United States, utilities warn that up to 20% of planned data center projects already face delays because of electricity transmission bottlenecks (IEA, 2025). Similar risks are emerging in Europe and Asia.

Since AI infrastructure depends disproportionately on low-cost, (and in the future low-carbon), and reliable electricity, only a handful of jurisdictions will attract investment on a large scale. This spatial inequality risks excluding much of the Global South, particularly Africa, from capturing digital value chains. Moreover, as electricity demand rises across sectors—transport electrification, industrial decarbonization, household demand—AI will increasingly become a direct competitor for scarce clean-power resources.

V. POWERING AI: THE SUPPLY-SIDE CHALLENGE

Having established the scale and trajectory of AI electricity demand, the next critical questions are where this power will come from, how it will be procured, and at what cost to energy systems and climate goals. Electricity is not a neutral input: its source, availability, and carbon intensity will determine whether AI becomes a driver of decarbonization or a new source of emissions lock-in. This section explores the current supply mix, the capacity of systems to match future growth, the evolving role of corporate procurement, and the structural trade-offs shaping AI infrastructure.

Current Supply and Regional Contrasts

Today, most electricity feeding AI and data-center infrastructure remains carbon-intensive. According to the IEA (2025), the physical mix of electricity consumed by data centers, including both onsite-generated and grid-supplied energy, is roughly 30% coal, 27% renewable energy, 26% natural gas, and 15% nuclear energy. Yet, this global figure hides sharp regional contrasts that define the emissions footprint and competitiveness of AI infrastructure.

In the United States, abundant domestic gas production, competitive pricing, and flexible generation make natural gas the dominant source, meeting over 40% of data center demand. Other factors supporting the use of natural gas include dispatchability, investment certainty, and project delivery speed. Renewables, primarily solar and wind, contribute about 24%, followed by coal (20%) and nuclear (15%) (IEA, 2025). In China, by contrast, coal accounts for about 70% of electricity supply to data centers, reflecting the east coast's heavy dependence on carbon-intensive grids, despite large-scale renewable buildout in the west. Coal is followed by renewables with nearly 20%, nuclear close to 10% and a marginal role for natural gas (IEA, 2025). Europe exhibits a more diverse supply mix, supported by larger shares of nuclear and renewables in countries including France and Sweden. However, systemic integration challenges and capacity constraints, especially in Ireland and the Netherlands, limit further progress, and occasionally reintroduce fossil backup.

These contrasts shape not only the emissions profile of AI infrastructure but also its cost structure, siting decisions, and vulnerability to policy pressure.

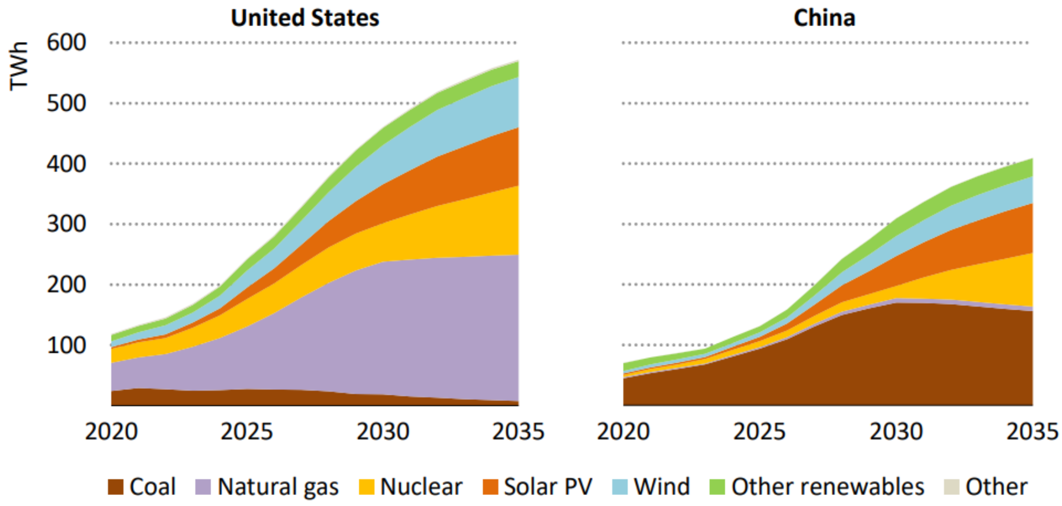
Can Supply Match the Growth Curve?

Regional disparities will only grow more significant as AI's electricity needs expand. Meeting the electricity demand projected in the IEA's base scenario, while remaining aligned with decarbonization goals, will require a rapid scaling of low-carbon generation, yet most systems are still adding fossil-fired capacity to ensure reliability.

In the United States, utilities are revising integrated resource plans to add new gas-fired capacity specifically for data centers, with projects such as Entergy Louisiana's 2 GW expansion for Meta exemplifying the trend (Figure 4). Similarly, NextEra Energy and GE Vernova are pursuing gas projects tailored to data center demand. To bring down emissions, some data center operators are considering fitting natural gas-fired plants with carbon capture in the long run (Skidmore, 2025). China faces a parallel dynamic with coal: Between 2024 and 2030, it is projected that coal and renewables will add roughly 90 TWh each to the electricity supply for data centers. Yet, given the heavier initial base, coal will maintain structural dominance through 2030.

Figure 4

Electricity Generation for Data Centers in the U.S. and China, IEA Base Scenario, 2020-2035



Source: IEA, 2025.

Europe is somewhat of an outlier, expected to cover 85% of incremental demand through renewables and nuclear. Yet public resistance, permitting delays, and grid congestion make rapid progress uncertain. Elsewhere, Japan and Korea are set to raise the shares of renewables and nuclear from 35% to nearly 60% by 2030, while Southeast Asia and India face slower transitions, with coal still playing a central role through the mid-2030s (IEA, 2025). These trajectories show that without accelerated investment in low-carbon baseload and storage, AI demand growth risks reinforcing fossil-fuel dependency, even in countries with ambitious climate goals.

Furthermore, given that most hyperscale capacity is clustered in the United States and China, it is plausible that the global uptake of AI services concentrates much of the associated energy and emissions burden in these jurisdictions. This suggests that AI adoption in other countries may indirectly add pressure to U.S. and Chinese grids, reinforcing the asymmetry between users of AI and hosts of its infrastructure.

Corporate Procurement and its Mismatch With Physical Supply

Alongside system-level supply, the strategies of hyperscalers themselves are shaping AI's real carbon footprint. Corporate renewable procurement has expanded rapidly, with over 30 GW of power purchase agreements (PPAs) signed globally in 2022. Yet most rely on 'annual matching', which balances yearly consumption against renewable certificates without aligning generation to hourly demand. This mismatch is critical. Dispatchable sources of electricity generation, including hydro, geothermal, and nuclear, can generally match hourly demand throughout the year, but this is not the case with variable renewables. In hourly terms, solar PV alone typically covers just 35%–45% of a data center's demand. As a result, during hours with low renewable output, most data centers still rely on grid electricity, often from fossil-based sources. This means data centers often rely on fossil-based grid power during hours of low renewable output, even while claiming '100% renewable' status. For instance, Microsoft's data centers in Ireland match their annual consumption with renewable PPAs, but during calm periods when wind generation falls, they must still draw on a gas-dependent grid (Grace & Cassidy, 2025).

Some hyperscalers are beginning to go further. For instance, Google's 478 MW offshore wind power purchase agreement allowed it to cover around 90% of its hourly consumption with clean energy in one location. Still, the remaining 10% of hours—when renewable generation dips—are covered by grid electricity that includes fossil-based (McKinsey, 2024). These experiments in "24/7 hourly matching," piloted in Denmark and Chile, show both the progress and the limits of current practice. Hybrid projects that combine solar, wind, and storage can raise hourly coverage from 40%–65% to as high as 90%, but at significantly higher cost and complexity. This underscores a central tension: green procurement on paper does not guarantee decarbonized operations in practice, and closing this gap will require regulatory oversight, new market instruments, and advances in storage economics.

Time, Cost, and Carbon: The New Trilemma of Data Center Location

Taken together, these supply patterns and procurement strategies crystallize into a new trilemma shaping the geography of AI infrastructure: balancing deployment speed, cost efficiency, and carbon intensity.

- **Deployment speed:** only solar PV and gas turbines align with typical data center construction timelines (one to two years). Wind, geothermal, and small modular reactors (a new generation of nuclear reactors designed for faster, modular deployment) generally require five to ten years, with nuclear and hydropower taking even longer.
- **Cost structure:** wind and solar are globally cost-competitive but require firming strategies to ensure reliability. Gas turbines remain attractive in low-carbon price environments because of their capital efficiency and dispatchability.
- **Carbon intensity:** coal and oil have the highest emissions per kWh. Gas reduces this by about

half. Renewables and nuclear offer zero direct emissions, but renewables in particular face integration challenges: their output depends on the sun and wind, which do not always match demand, requiring backup generation, storage, or new grid infrastructure. Without these measures, the real-world carbon savings of renewables can be diluted.

This trilemma is already influencing siting choices: Norway offers low-carbon, low-cost electricity but limited scale; the U.S. Gulf Coast offers speed and capacity but at the cost of higher emissions; Kenya offers geothermal potential but faces infrastructure bottlenecks. The geography of AI will therefore be defined as much by energy constraints as by digital innovation capacity.

VI. THE GEOPOLITICS OF AI: SOVEREIGNTY AND STRATEGIC VULNERABILITIES

Having examined how AI's rising electricity demand will be met, it becomes clear that the challenge is not confined to generation technologies or procurement models. The scaling of AI also creates new geopolitical dependencies: on grids capable of absorbing hyperscale loads, on specialized equipment such as transformers and turbines, on semiconductor supply chains concentrated in East Asia, and on critical minerals subject to export controls and strategic rivalry. In other words, the debate around 'powering AI' is inseparable from questions of sovereignty, resilience, and vulnerability in global supply chains. This section unpacks those dynamics in terms of the sovereignty of electricity infrastructure, and the strategic control of compute power.

6.1. Electricity as Geopolitical Infrastructure

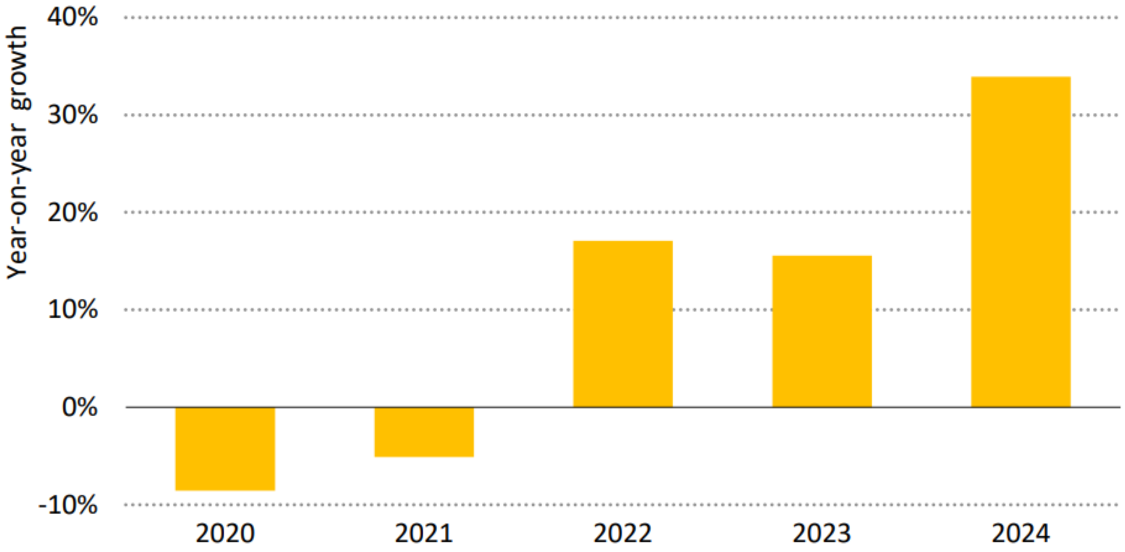
As detailed in section IV, the geography of AI infrastructure is being redrawn around access to abundant, reliable, and increasingly low-carbon electricity. The shift is not just about cost optimization. It reflects a deeper power asymmetry: countries with mature grids, fast interconnection permitting, and surplus capacity are increasingly monopolizing AI-related investment. These dynamics risk marginalizing nations with underdeveloped, fragile, or slow to modernize power systems, which cannot meet the operational and locational demands of hyperscale computing.

The emergence of private power foundries in the United States offers a stark example. Chevron and Engine No.1 are developing dedicated 4 GW gas-fired facilities to directly power data centers, using GE Vernova turbines, bypassing public grid systems altogether. While such arrangements offer predictability and scale, they mark a shift toward the privatization of energy allocation, undermining grid coordination, public oversight, and equitable access. In constrained markets such as Ireland, Virginia, and the Netherlands, grid saturation has already led to multiyear delays in new interconnections. These bottlenecks are no longer technical, they reflect institutional lag, governance inertia, and fragmented planning authority, which now act as geopolitical constraints (IEA, 2025).

This mismatch between the AI sector's rapid growth and inertia of energy infrastructure is particularly acute in hardware supply chains (Figure 5). According to the IEA (2025), global prices for transformers have more than doubled since 2018, while lead times for large units now exceed 30 months. The price of grain-oriented electrical steel (GOES), which constitutes up to 20% of a transformer's cost, has nearly doubled since 2021. Global cable prices have also surged, driven by rising copper and aluminum costs, strong demand from electrification and digitalization, and chronic underinvestment in manufacturing capacity.

Figure 5

Increase in Power Transformer Order Backlog, Selected Manufacturing Companies, 2020-2024



Source: IEA, 2025. Based on order backlogs of Hitachi Energy, Schneider Electric, Siemens Energy, GE Vernova.

These bottlenecks are not randomly distributed. Between 2018 and 2023, global trade in power transformers rose by 80%, with China, Türkiye, Korea, and Italy accounting for over half of global supply. Meanwhile, the U.S. and the EU doubled their imports, underscoring how even advanced economies depend on international suppliers for their core digital infrastructure. The lack of local manufacturing capacity, especially for high-specification transmission-grade components, constitutes a growing strategic liability.

In this context, energy sovereignty must be redefined. It is no longer merely about generating capacity or decarbonization targets. It now includes the ability to deploy, interconnect, and maintain the physical systems, transformers, cables, substations, and grid controls, on which compute infrastructure depends. Without this capacity, even nations rich in data, AI talent, or capital may find themselves excluded from the next phase of digital-industrial development.

6.2. Chips, Lithography, and Critical Minerals as Strategic Bottlenecks

If electricity is the lifeblood of AI, semiconductors are its central nervous system. The production of these chips, especially GPUs used to train and deploy large AI models, relies on hyper-specialized supply chains characterized by extreme geographic concentration, long ramp-up times, and strategic control.

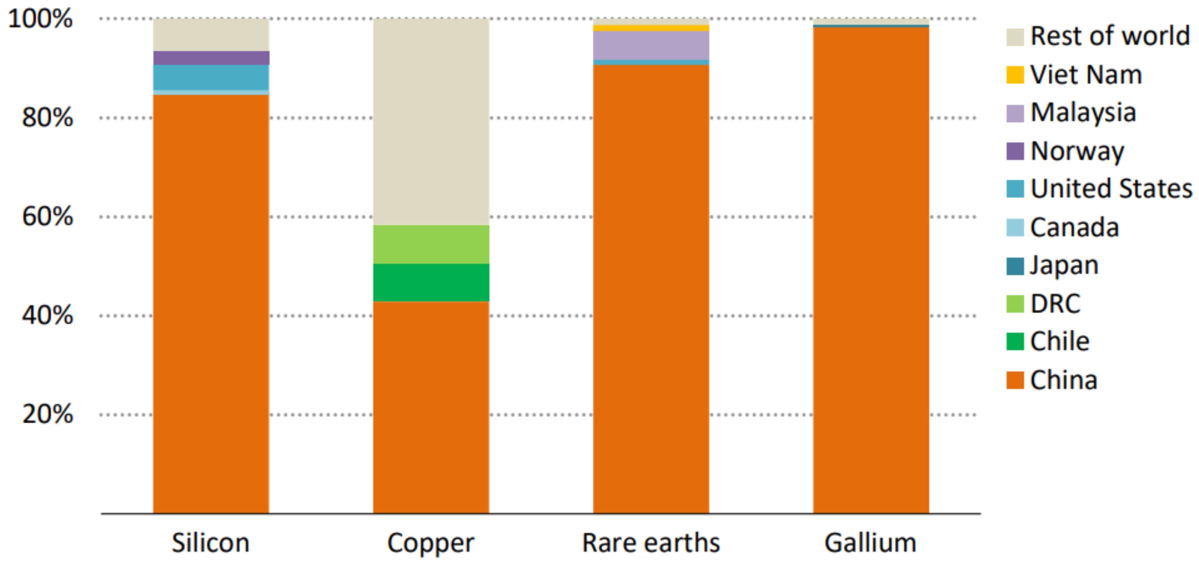
Chip design remains dominated by U.S. firms such as NVIDIA, AMD, and Intel. Yet, over 70% of advanced semiconductor manufacturing is outsourced to East Asia, primarily Taiwan's TSMC and South Korea's Samsung Foundry (IEA, 2025). This division of labor between Western intellectual property and Asian fabrication capacity creates deep-seated vulnerabilities. A geopolitical shock, whether conflict in the Taiwan Strait or export restrictions, could instantly sever global access to frontier compute capabilities.

This fragility is exacerbated by the monopoly of ASML, a Dutch company that is the world's sole supplier of extreme ultraviolet (EUV) lithography machines — indispensable for producing chips at the 5-nanometer scale and below (Tarasov, 2022, IEA, 2025). These machines, priced at $183 million to $380 million, are thus critical to advanced semiconductor manufacturing. In 2024, only 44 EUV units were delivered worldwide. Each machine requires over 100,000 components, an 18-month assembly period, and coordination across thousands of suppliers. ASML, is therefore, not merely a supplier but a strategic bottleneck in global digital sovereignty.

Yet, even chips require raw materials (Figure 6). AI hardware relies on ultrapure silicon, copper, gallium, rare earth elements, and indium. According to IEA (2025), data centers could account for 2% of global and silicon demand, 3% of rare earth consumption, and a staggering 11% of gallium use by 2030. These materials are finite, and their supply chains are geopolitically fraught. China, for example, dominates the refining and export of gallium and other high-performance metals. In 2023-2024, it implemented successive export controls on gallium, germanium, graphite, tungsten, and bismuth in retaliation to U.S. technology sanctions. The effect was immediate: gallium prices outside China doubled in six months.

Figure 6

Geographical Concentration of the Supply of Selected Refined Critical Minerals Needed for Data Center Expansion, 2024



Source: IEA, 2025.

Meanwhile, the U.S. has imposed escalating export controls on AI chips, including NVIDIA's A100 and H100, effectively cutting Chinese buyers off from cutting-edge hardware. This reveals a structural asymmetry in the technological contest: while China holds leverage over upstream raw materials and refining, the U.S. and its allies retain control over the most sophisticated nodes of semiconductor design, lithography, and fabrication, concentrated in firms such as ASML, TSMC, and NVIDIA. Both sides are interdependent, but the chokepoints embedded in advanced manufacturing ecosystems cannot be replicated rapidly. Despite significant state-led investment, replicating TSMC's node leadership or ASML's lithography capacity remains a decade-scale challenge for China.

While the semiconductor value chain is dominated by the United States, East Asia, the Netherlands, and China, other regions also play roles that are often overlooked. As figure 6 shows, Chile and the Democratic Republic of Congo play important roles in global copper supply, while smaller shares of silicon and rare earths come from the United States, Canada, and Malaysia. Beyond these, Australia is a major producer of lithium, the DRC provides most of the world's cobalt, and South Africa leads in platinum-group metals. Yet these roles remain largely confined to extraction, whereas the higher-value segments of design, lithography, and fabrication are concentrated in a handful of actors. This asymmetry underscores the broader sovereignty challenge: many countries are locked into supplier positions, while strategic control and value capture remain elsewhere.

What emerges is a triangular geopolitical dependency, between energy infrastructure, compute hardware, and access to strategic resources. This determines a country's ability to participate in the AI economy. These are not technical constraints, but structural limitations on national agency. Nations that cannot secure the full stack, from grid capacity to transformers, from GPUs to gallium, risk being relegated to the periphery of digital production.

VII. AFRICA IN THE AI–ENERGY NEXUS: RISKS AND STRATEGIC ENTRY POINTS

The geopolitics of AI outlined in Section VI, with digital sovereignty defined by access to energy and materials, finds its sharpest expression in Africa. The continent is indispensable to the material foundations of the AI economy, yet structurally constrained in its capacity to harness these technologies for its own development.

Africa has the bulk of global cobalt reserves, significant shares of manganese and rare earths, and growing importance in gallium and graphite markets—materials that are indispensable for batteries, semiconductors, and renewable technologies underpinning the AI–energy economy. At the same time, over 600 million Africans still lack access to reliable electricity (World Bank, 2023), grid outages remain chronic, and domestic compute capacity is negligible compared to global leaders. While the United States, China, and Europe consolidate their control over compute infrastructure and energy-intensive data systems, Africa risks being locked into a familiar role: supplying raw materials while depending on foreign actors for digital infrastructure, cloud services, and regulatory frameworks.

Infrastructure Readiness and Uneven Digital Geographic Starting Points

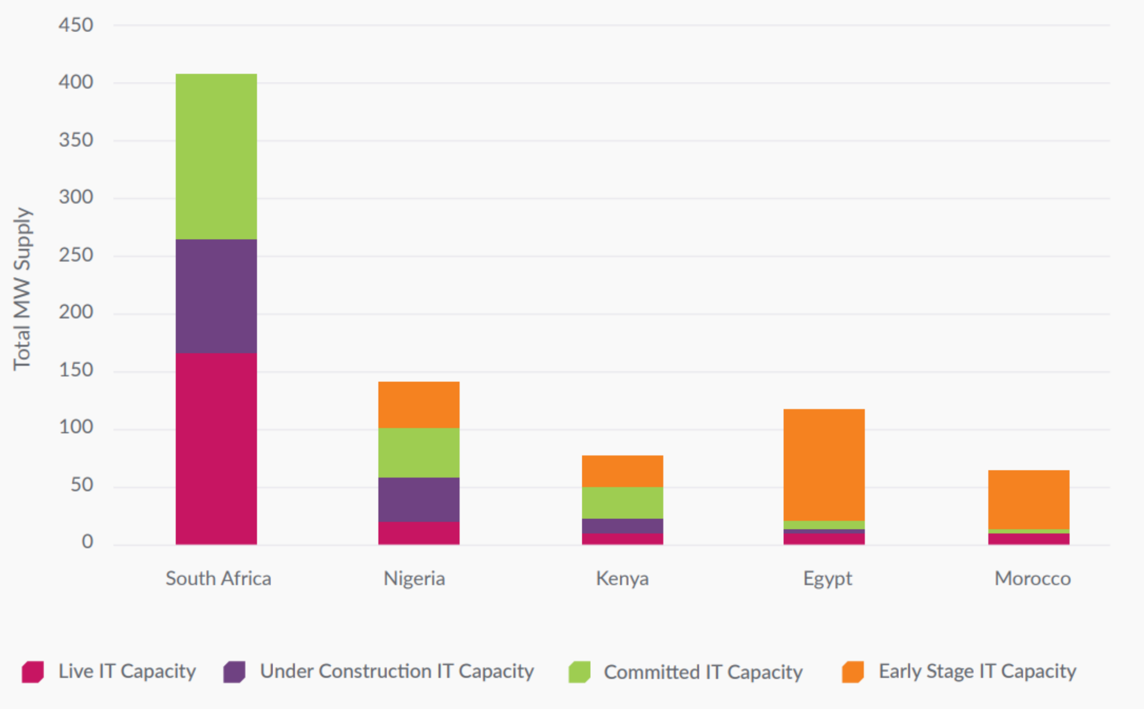
AI take-up in Africa is starting from a structurally uneven base. Only 47% of Africans have access to electricity, with stark contrasts between North Africa (>90%) and sub-Saharan Africa (<40%) (IEA, 2024). Grid reliability is a binding constraint: sub-Saharan countries face average annual outages exceeding 700 hours (World Bank, 2023), levels incompatible with AI-enabled grid management. Internet penetration averages 40%, but ranges from 65% in North Africa to below 20% in Central Africa (AU, 2020). These deficits risk confining AI deployment to isolated, largely metropolitan digital enclaves, bypassing broader developmental needs.

Within this constrained landscape, five national hubs dominate data center development: South Africa, Nigeria, Egypt, Kenya, and Morocco. Together they accounted for more than 800 MW of IT power as of 2023 (DCByte, 2024).

South Africa alone represents about 350 MW, making it the largest data center market in Africa, based on available critical IT load. It hosts around 40 colocation facilities built to the Tier III standard — shared data centers designed with redundant power and cooling systems that allow maintenance without downtime. Johannesburg and Cape Town also remain the only African cities with fully established hyperscale cloud regions from Amazon Web Services, Google, and Microsoft.

Figure 8

IT Capacity Breakdown by Country[3]



Source: DCByte, 2024.

Nigeria (140 MW) and Egypt (118 MW) are expanding rapidly, while Kenya (79 MW) positions itself as East Africa's digital hub. Morocco (65 MW), though smaller, leverages strategic submarine cable links to Europe. Beyond these headline figures, growth trends are notable: Kenya has recorded an 84% compound annual growth rate in IT capacity (2018–2022), Nigeria 50%, and Egypt 17%. These hubs reflect both demand drivers (population, financial services, telecom growth), and structural

[3] IT capacity (critical IT load): the maximum electrical power, measured in megawatts, that a data center can deliver to its computing equipment (servers, storage, and networking), excluding cooling and other auxiliary systems.

enablers (grid access, submarine cable connectivity, regulatory openness).

Yet the majority of announced capacity remains under construction, rather than live. This gap between planned and live facilities has several implications. First, it means that headline capacity figures may overstate the actual resources currently available to support AI and cloud services. Second, construction delays — often linked to financing, permitting, or power access — highlight the fragility of the ecosystem, where grid bottlenecks and regulatory hurdles can stall deployment. Third, it underscores a temporal mismatch: demand for digital services is growing rapidly, but supply will only materialize gradually, creating near-term risks of congestion and service concentration in a handful of operational hubs. Finally, reliance on under-construction projects reveals the uncertainty of Africa's data center trajectory: unless projects are delivered on schedule and grid-integrated, much of the projected digital capacity may remain aspirational rather than transformative. Therefore, without parallel investment in regional interconnections and backbone infrastructure, the African digital divide could widen.

Sovereignty in Compute and Data: Dependency Risks

Africa's emerging digital infrastructure is overwhelmingly under foreign ownership and control, raising structural questions of sovereignty. Hyperscale facilities across the continent are largely operated by Microsoft Azure, Amazon Web Services, Google Cloud, and Huawei Cloud, often in joint ventures with local telecom operators. While these partnerships expand capacity, they externalize control over core layers of the digital economy—data storage, processing power, and cloud governance. Even when AI applications are deployed domestically, such as predictive maintenance for Eskom in South Africa or grid pilots in Egypt, the supporting compute infrastructure is designed, operated, and governed offshore.

This dependency creates two major sovereignty risks. First, access to compute power is a strategic vulnerability. Export controls on advanced GPUs demonstrate how quickly supply can become a geopolitical lever, leaving African states with no fallback capacity. Second, data governance is externalized. Hyperscale operators determine where data is stored, how it is processed, and under which jurisdictional rules—constraining governments' ability to manage energy-related data, a resource that is increasingly strategic for forecasting, grid stability, and national security.

The structural analogy with Africa's industrial past is clear. The continent provides critical raw materials—cobalt, manganese, platinum, rare earths—yet captures little of the downstream value. For instance, the Democratic Republic of Congo produces over 70% of global cobalt but captures less than 10% of the value of battery manufacturing (OECD, 2023). In the AI economy, the same pattern risks repeating: Africa supplies critical raw materials — such as cobalt, manganese, platinum-group metals, and rare earth elements — that are indispensable for manufacturing semiconductors and powering data centers, but it lacks domestic chip fabrication plants or large-scale facilities to process these resources, thus remaining dependent on importing high-value compute and cloud services at high cost.

The African Union's Digital Transformation Strategy (2020–2030) acknowledges this risk and calls for building regional capacity in data hosting and processing. Progress, however, remains slow and uneven. Without deliberate policy and investment, Africa's position in the AI–energy nexus may replicate its historic role in commodities: critical yet peripheral, subject to "digital recolonization" rather than digital sovereignty (Azeroual, 2024).

Human Capital and Regulatory Gaps

The deployment of AI in energy is not only about infrastructure and minerals; it is about people and governance. Currently, fewer than 3% of African tertiary graduates specialize in engineering, ICT, or energy-related fields (World Bank, 2024). Research ecosystems remain thinly spread, with isolated centers such as the African Institute for Mathematical Sciences (AIMS), and the University of Cape Town's AI Lab. Without deliberate investment in research and training ecosystems that integrate universities, utilities, and industry, Africa risks remaining a technology taker rather than a co-developer.

Furthermore, Africa's regulatory frameworks are weak. Only a handful of countries (South Africa, Kenya, Tunisia) have begun to establish regulatory sandboxes for AI in energy. In most jurisdictions, utility incentives remain misaligned, rewarding higher electricity sales (revenue expansion) rather than cost reduction or efficiency gains from AI. Furthermore, the global challenge of trust and explainability in AI is magnified in Africa, where fragile power systems leave little margin for error. A 'leapfrogging' approach—deploying AI directly in real-time grid management—could risk destabilization. More pragmatic sequencing would focus first on non-critical functions such as forecasting, asset management, and planning, before moving toward system-critical operations.

Strategic Pathways: Regional Integration, Industrial Policy, and Governance

Africa's engagement with the AI-energy nexus must be framed as a question of sovereignty and strategic choice, not merely of technology transfer. Three policy priorities stand out:

1. **Regional Infrastructure Integration:** Africa's reliance on fragmented national projects limits both bargaining power and system resilience. AfCFTA and regional power pools provide potential vehicles for collective scale, but their current limitations are significant. AfCFTA's digital trade protocols remain embryonic, and power pools are constrained by underdeveloped transmission links and uneven regulatory frameworks. Leveraging these platforms therefore requires more than rhetorical alignment: it calls for investment in cross-border grids, regulatory convergence, and shared digital backbone infrastructure. Without such measures, these regional frameworks risk remaining aspirational, leaving individual states exposed to asymmetric dependencies with global hyperscalers.
2. **Mineral-to-Industrial Value Chains:** To avoid remaining locked at the resource-supplier stage, Africa must translate its mineral endowments into industrial capacity directly relevant to the AI economy. This requires linking cobalt, platinum-group metals, and rare earth extraction to domestic refining, battery and component manufacturing, and eventually AI-related hardware assembly. The challenge is scale: no single national market can sustain competitive fabrication or component industries. Regional approaches are therefore essential — pooling demand through AfCFTA, coordinating incentives for local processing, and attracting joint ventures with equipment makers. Current strategies often stop at basic beneficiation, but without deeper integration into global semiconductor and data center supply chains, Africa's mineral wealth will continue to generate rents without structural transformation.
3. **Human Capital and Governance:** The ability to shape AI deployment around Africa's development requires long-term investment in specialized training for engineers, data scientists, and regulators who can adapt global technologies to local contexts. Regulatory sandboxes can provide controlled environments for experimentation, balancing innovation with safeguards

for energy security, privacy, and equity. Beyond isolated initiatives, centers of excellence — whether regional hubs for AI in energy systems or cross-border research networks — can anchor talent, attract investment, and set locally relevant standards. In the short term, African countries may need to rely on AI-as-a-Service through global providers, while gradually developing local capacity for data, compute, and energy systems.

The alternative is stark: a fragmented, donor-driven deployment of AI in energy, leaving Africa as a peripheral actor in the intelligence economy, supplying minerals but excluded from value creation. There is an opportunity to avoid this outcome, but it requires deliberate choices today in relation to infrastructure, industrial strategy, and governance.

CONCLUSION

AI has moved from a niche technological innovation to a systemic driver of economic and energy dynamics. Its rapid expansion underscores a dual reality: AI is both a tool for optimizing energy systems, and a major source of new electricity demand. This paradox is reshaping industrial strategies, infrastructure planning, and geopolitical hierarchies. The evidence presented in this paper demonstrates that AI growth demands more energy, infrastructure, and critical materials, and will increasingly test the resilience and adaptability of national energy systems.

For advanced economies, the policy challenge lies in balancing the acceleration of AI adoption with commitments to decarbonization and grid adequacy. The rise of hyperscale data centers, clustered in a handful of jurisdictions, will exacerbate congestion, reinforce spatial inequality, and force trade-offs between clean-energy allocation to digital infrastructure, versus other strategic uses, such as transport electrification or industrial decarbonization. The new trilemma of time, cost, and carbon in data center siting means that energy policy, innovation policy, and access to strategic resources will together determine the geography of AI leadership.

For emerging markets, the stakes are different but no less profound. The concentration of compute power, lithography, and semiconductor supply chains in a few economies creates strategic dependencies that mirror past patterns of industrial asymmetry. Energy sovereignty today must be redefined: it extends beyond generation capacity to include transformers, grids, compute infrastructure, and the ability to govern the data flows that underpin AI. If countries are unable to secure this full stack, they risk being left as passive users of imported intelligence, rather than co-creators of the digital-industrial economy.

Africa illustrates these dynamics with particular clarity. The continent holds significant reserves of minerals essential for powering the electricity systems that underpin AI, yet remains structurally excluded from higher-value segments of the compute and data chain. Digital infrastructure is growing, but in narrow enclaves centered on a handful of hubs, with systemic deficits in grid reliability, internet penetration, and human capital. Without deliberate policies, Africa risks repeating a familiar pattern: exporting raw materials while importing value-added technologies at high cost. To avoid this trajectory, the continent must leverage its resource position to move beyond raw material supply toward processing and manufacturing, while investing in integrated regional data and electricity infrastructure. Addressing this gap also requires solutions adapted to African contexts, such as smaller-scale modular data centers, renewable-powered edge computing, and regional cloud collaborations, alongside strong regulatory frameworks and human capital development to enable sovereign participation in the AI–energy nexus.

While AI and energy remain distinct policy domains, their points of intersection are becoming increasingly strategic. Managing this intersection will determine whether AI becomes an accelerator of the global energy transition or an additional source of systemic strain. Countries that align digital policy with energy planning, and that treat compute infrastructure as a strategic asset rather than a neutral input, will be better positioned to capture value, mitigate vulnerabilities, and ensure a balanced trajectory toward net-zero emissions. Conversely, those that fail to adapt risk deepening asymmetries, ceding sovereignty, and missing one of the defining technological shifts of the twenty-first century.

The choice is therefore not about whether AI will transform energy systems—it already is—but about who will shape, govern, and benefit from this transformation.

References

- Adjei, K, Y. et al. (2025), Optimizing well placement and reducing costs using AI-driven automation in drilling operations, World Journal of Advanced Research and Reviews, 2025, 25(02), 1029-1038, https://doi.org/10.30574/wjarr.2025.25.2.0436

- Azaroual, F. (2024), L'Intelligence Artificielle en Afrique : défis et opportunités, Policy Brief, https://www.policycenter.ma/publications/lintelligence-artificielle-en-afrique-defis-et-opportunites

- Araya-Polo, M. et al. (2017), Automated fault detection without seismic processing, The Leading Edge, pp. 208-214, https://doi.org/10.1190/tle36030208.1

- Cerebras (2022), https://www.cerebras.ai/customer-spotlights/totalenergies Clarke, P. et al. (2025), A new dotcom bubble? AI hype has yet to translate into profits, Science & Technology, The Observer, https://observer.co.uk/news/science-technology/article/a-new-dotcom-bubble-ai-hype-has-yet-to-translate-into-profits

- DCByte (2024), Africa's Key Data Centre Markets, Market Spotlight, https://www.dcbyte.com/wp-content/uploads/2023/07/DC-BYTE_MARKET-SPOTLIGHT_AFRICA.pdf?utm_campaign=Market%20Spotlight%202024&utm_medium=email&_hsenc=p2ANqtz-92MMEMJ_iirBQwbxX7m-oWN9OvkLxeR5jQbXGWpuNfvk36dE4iun4PoeMQ3Y5wx6MUMp_ZJYgvvpg8UWdV5QITMlnsEA&_hsmi=324969182&utm_content=324969182&utm_source=hs_automation

- EIA (2024), How much electricity does an American home use? https://www.eia.gov/tools/faqs/faq.php?id=97&t=3

- Enel (2022), Enel and Myst AI: Optimizing energy forecasts, https://openinnovability.enel.com/stories/articles/2022/04/mystai-powering-sustainableai

- EpochAI (2024), Notable AI Models, https://epoch.ai/data/notable-ai-models#data-insights

- EpochAI (2025), How Much Energy Does CHatGPT Use? https://epoch.ai/gradient-updates/how-much-energy-does-chatgpt-use

- Fulton, J. et al. (2024), Forecasting regional PV power in Great-Britain with a multi-model late fusion network, https://s3.us-east-1.amazonaws.com/climate-change-ai/papers/iclr2024/46/paper.pdf

- Garcia Bardon, M., et al. (2021), DTCO including Sustainability: Power-Performance-AreaCost Environmental score (PPACE) Analysis for Logic Technologies, IEEE, https://doi.org/10.1109/IEDM13553.2020.9372004

- GE Vernova (2022), GE Using AI/ML to Reduce Wind Turbine Logistics and Installation Costs, https://www.gevernova.com/news/press-releases/ge-using-aiml-to-reduce-windturbine-logistics-and-installation-costs

- Google DeepMind (2019), Machine learning can boost the value of wind energy, https://deepmind.google/discover/blog/machine-learning-can-boost-the-value-of-windenergy/

- Grace, M., et al. (2025), Data Centres in Ireland – Energy Concerns, Mayson Hayes & Curran, https://www.mhc.ie/latest/insights/data-centres-in-ireland-energy-concerns

- Greenpeace (2023), Invisible Emissions, https://www.greenpeace.org/eastasia/invisible-emissions/

- Hitachi Energy (2024), Nostradamus® AI Energy Forecasting Software, https://www.hitachienergy.com/products-and-solutions/energy-portfoliomanagement/market-analysis/nostradamus-ai-energy-forecasting-software

- Iberdrola (2016), MeteoFlow Project, https://www.iberdrola.com/innovation/meteoflow-project

- IEA. (2025), Energy and AI, Report, https://www.iea.org/reports/energy-and-ai

- JPT (Journal of Petroleum Technology) (2022), Chevron Work flow Reinforces Importance of Simulation to Predictive Behaviors, https://jpt.spe.org/chevron-workflow-reinforcesimportance-of-simulation-to-predictive-behaviors

- Kuang, L. et al. (2021), Application and development trend of artificial intelligence in petroleum exploration and development, Petroleum Exploration and Development, https://doi.org/10.1016/S1876-3804(21)60001-0

- Mckinsey (2024), How hyperscalers are fueling the race for 24/7 clean power, https://www.mckinsey.com/industries/electric-power-and-natural-gas/our-insights/how-hyperscalers-are-fueling-the-race-for-24-7-clean-power#/

- PNNL (Pacific Northwest National Laboratory) (2024), Faster, More Informed Environmental Permitting with AI-Guided Support, https://www.pnnl.gov/newsmedia/faster-more-informed-environmental-permitting-ai-guided-support

- Skidmore, Z. (2025), NextEra partners with GE Vernova on gas for data centers, will restart Iowa nuclear plant, Data Center Dynamics, https://www.datacenterdynamics.com/en/news/nextera-partners-with-ge-vernova-on-gas-for-data-centers-will-restart-iowa-nuclear-plant/

- Tarasov, K. (2022), ASML is the only company making the $200 million machines needed to print every advanced microchip. Here's an inside look, CNBC, https://www.cnbc.com/2022/03/23/inside-asml-the-company-advanced-chipmakers-use-for-euv-lithography.html

- TotalEnergies (2022), Papua New Guinea: TotalEnergies Announces New Milestone towards Papua LNG Development, Press Release, https://totalenergies.com/sites/g/files/nytnzq121/files/documents/2022-07/EN_Papua_New_Guinea_TotalEnergies_Announces_New_Milestone_Papua_LNG_Development.pdf

- Xia H., Strayer A. and Ravikumar A.P. (2024), The role of emission size distribution on the efficacy of new technologies to reduce methane emissions from the oil and gas sector, Environmental Science and Technology, pp. 1088 – 1096, https://pubs.acs.org/doi/10.1021/acs.est.3c05245

ABOUT THE AUTHOR

RIM BERAHAB

Rim Berahab is Senior Economist at the Policy Center for the New South, which she joined in 2014. She is currently working on themes related to energy issues and their impacts on economic growth and long-term development. Her research areas also cover trade and regional integration challenges in Africa. Previously, she has also worked on questions related to gender inequalities in the labor market of North African countries. Rim spent three months at the International Monetary Fund (IMF), in 2016, in the Commodities Unit of the Research Department. She holds a State Engineering degree from the National Institute of Statistics and Applied Economics (INSEA).

ABOUT THE POLICY CENTER FOR THE NEW SOUTH

The Policy Center for the New South (PCNS) is a Moroccan think tank aiming to contribute to the improvement of economic and social public policies that challenge Morocco and the rest of Africa as integral parts of the global South.

The PCNS pleads for an open, accountable and enterprising "new South" that defines its own narratives and mental maps around the Mediterranean and South Atlantic basins, as part of a forward-looking relationship with the rest of the world. Through its analytical endeavours, the think tank aims to support the development of public policies in Africa and to give the floor to experts from the South. This stance is focused on dialogue and partnership, and aims to cultivate African expertise and excellence needed for the accurate analysis of African and global challenges and the suggestion of appropriate solutions.

All opinions expressed in this publication are those of the author.

Policy Center for the New South

Rabat Campus of Mohammed VI Polytechnic University,
Rocade Rabat Salé - 11103
Email : contact@policycenter.ma
Phone : +212 (0) 537 54 04 04
Fax : +212 (0) 537 71 31 54

www.policycenter.ma

POLICY CENTER
FOR THE NEW SOUTH

RePEc