Energy Sovereignty: The Layer Nothing Can Compensate

Volume 4 and closure of the 'Digital Sovereignty' series. Among the four layers identified in Volume 1, three compose by architectural instrument. The fourth is requested from another. It does not merely shift the AI decision into long time horizons; it brings AI into the strategic arbitrage matrix of the energy state.

Take-Away

- "Volume 4 establishes the structural asymmetry of Europe's energy layer for AI: among four sovereignty layers (infrastructure, fabric, model, energy), only the energy layer cannot be composed by architectural instrument. It is allocated by the Transmission System Operator within multisectoral planning, not designed by the deployer."
- "IEA reports global datacenter electricity consumption rose from 415 TWh in 2024 to 485 TWh in 2025 (+17%), projected to 945 TWh by 2030. ENTSO-E warns on 8 May 2026 that European TSOs may have to reduce renewable penetration if datacenter growth remains unmanaged. RTE connection delays in France: 18 to 36 months for projects above 10 MW."
- "A nuclear Power Purchase Agreement commits 20 years (Microsoft-Constellation 835 MW, Three Mile Island restart 2027; Amazon-Talen 1,920 MW until 2042; Meta-Constellation 1,121 MW from June 2027). A frontier model lifecycle is 12 to 18 months. The deployer signs an energy horizon outliving every model trained on it."
- "US deployers internalize generation: Stargate $500B/10 GW by 2029, Anthropic $50B in November 2025, $400B combined hyperscaler capex in 2025, 38 GW captive gas in development. European deployers compose with the regulated grid: Mistral 1.4 GW campus near Melun for 2028, AWS Sovereign Cloud Brandenburg 15 January 2026, EDF EPR2 €72.8B program."
- "GPAI Code of Practice published 10 July 2025 mandates training-energy disclosure (Article 53 AI Act, Annex XI), but exempts providers lacking compute-side data, signaling causal non-localizability. CSRD applies; Empowering Consumers for the Green Transition Directive applies 27 September 2026; Green Claims proposal dormant since 20 June 2025."

- "The fourth doctrinal operator is allocation, not proof, diversification, or instrumentation. Sovereignty under allocation: the deployer negotiates with the TSO and the planning state. The series closes on a quadripartite composite doctrine and opens a passage to executable audit instruments opposable to architecture review boards."

keywords

- energy sovereignty digital sovereignty AI infrastructure
- datacenters RTE SDDR EPR2 nuclear PPA
- GPAI Code of Practice NIS2 CSRD ECGT
- Three Mile Island Stargate ENTSO-E
- AWS Sovereign Cloud series: digital-sovereignty seriesIndex: 4 seriesTitle: "Digital Sovereignty" relatedTo:
- architectural-sovereignty sovereignty-stack-not-label model-sovereignty
- eda-essential-complement-agentic-ai hexagonal-architecture-governability
- ai-governance-architecture

Recap and frame

Volume 1 established that digital sovereignty is not a political debate but a condition for capitalizing performance, signed by a chain of four ports: identity, location, custody, registry. Volume 2 demonstrated that this sovereignty is a stack, not a label, and that it reads layer by layer across the seven floors of the material fabric. Volume 3 established that model sovereignty is not read in the publisher's passport but in a triad of evidentiary chains: inspectable weights, documented data, governed versions. The closing echo-triplet of Volume 3 stated the three-layer composite doctrine: *a SecNumCloud certification immunizes the operator, not the silicon it operates; the silicon operates a model, not the auditability chain that makes it defensible; a model that cannot be frozen is a model that cannot be certified*.

What remains is the fourth layer, flagged from Volume 1 as the authentic limit of the doctrine. The energy layer does not sign by cryptography, does not compose by supplier diversification, does not instrument by contractual triad. It is allocated. The present volume treats this fourth layer and closes the four-volume arc.

The domain of validity is unchanged: GDPR, HDS, NIS2, high-risk AI Act, MDR, IVDR. Volume 4 adds three regulatory corpora that have become structuring to varying degrees:

the CSRD on sustainability reporting, already applicable; the Empowering Consumers for the Green Transition (ECGT) Directive on environmental claims to consumers, applicable from 27 September 2026; and NIS2 in its digital-infrastructure scope, under which large-enough datacenters are already directly qualified.

The distinction that cuts, and the collision it reveals

The opening gap to draw is elementary. *Installed capacity is a commercial category, allocated capacity is an engineering category*. The whole doctrine of this volume rests in this distinction, and in the collision it reveals between two regimes of reality.

On 19 May 2025, at the Choose France summit in Versailles, Bpifrance, MGX, Mistral AI, and NVIDIA announced a joint venture to develop in the Paris region an AI campus targeting 1.4 gigawatts, roughly 85% of the electrical output of an EPR such as Flamanville 3. Construction is planned for the second half of 2026, commercial operation announced for 2028, first investment tranche €8.5 billion. This announcement engages the investment decision. It does not guarantee the grid decision. The Réseau de Transport d'Électricité (RTE) published its 2025 Schéma Décennal de Développement du Réseau (SDDR) and submitted it to public debate under the auspices of the Commission nationale du débat public from 4 September 2025 to 14 January 2026. The SDDR articulates three priorities to be arbitrated together: industrial decarbonization, accommodation of new consumers (datacenters included), and development of decarbonized generation.

The gap between the announcement and the connection is not a calendar detail. It is the expression of a collision. *The narrative economy of AI operates on financial objects; the electrical economy operates on physical objects.* The hyperscaler industrial storytelling lines up announced gigawatts, announced capex, announced campuses, announced GPUs. The electrical system runs on firm allocation, frequency stability, inertia, local congestion, transformers, dispatchability. The two registers do not contradict each other in theory; they do not align in practice. The nationality of the capital does not shorten the queue, and the political tone of a summit does not rewrite a decennial schema.

The predictable objection is that coordination exists precisely for this. True. But the doctrine does not require the absence of coordination. It requires reading the connection agreement rather than the press release, and holding the distinction between a commercial category and an engineering category.

Triad of energy constraints

As in Volume 3, the doctrine deploys in a triad. Three structuring constraints, all irreducible.

1. **First constraint: allocated capacity.** Installed capacity in the European fleet is not, on a 2030 horizon, the limiting factor. Local allocation is rivalrous. According to the IEA, global datacenter electricity consumption rose from 415 TWh in 2024 to 485 TWh in 2025 (+17%), with a base-case projection of 945 TWh by 2030, and European growth of roughly 70% over the period. ENTSO-E, in its report *Data centres and the power system* published 8 May 2026, counts about 12.7 GW of installed IT capacity across 10,500 European datacenters with at least 50 kW of IT load, and warns that transmission operators could be forced to reduce renewable penetration if datacenter growth is not regulated. The doctrine here is not that growth is excessive. It is that *allocation is rivalrous*, and that the SDDR's arbitrage criterion is not the local maximization of a digital actor but the multisectoral planning of the national electrical system.
2. **Second constraint: connection calendar.** RTE recommends, for any project above 10 MW, a delay of 18 to 36 months between request for study and commissioning. In the United States, the Lawrence Berkeley Laboratory documented in its *Queued Up* report of May 2025 that the typical duration from connection request to commercial operation rose from under two years for projects built in 2000 to 2007 to over four years for those built in 2018 to 2023, with a median of five years for the 2023 cohort. The doctrine here is not that it is long. It is that *it is longer than any AI product cycle the deployer commits to*. A frontier model generation amortizes over 12 to 18 months; an industrial connection is processed over 24 to 36; a structuring electrical generation program over 10 to 15 years. The asymmetry is structural, not cyclical.
3. **Third constraint: temporal concordance between load and generation.** An AI training load runs at 80 to 95% capacity factor, with large and rapid power transitions documented by the IEA in its *Key Questions on Energy and AI* report of 2026. A photovoltaic plant operates at an annual mean capacity factor of 12 to 15%, an onshore wind farm at 25 to 30%, an offshore wind farm at 35 to 45%. Concordance is not a physical contradiction so long as the system is compensated by storage, dispatchable backup, or interconnection. It becomes an arbitrage difficulty as soon as the grid operator must guarantee service to all consumers without degrading stability. The doctrine here is not that renewables are unsuited to AI. It is that *the capacity factor is proven, not promised*. A claim of 100% renewable supply is verified in the hourly mix, in time-stamped guarantees of origin, in the supply contract, or it is not verified at all.

Temporal asymmetry and matrix shift

It is here that the energy layer most sharply distinguishes itself from the previous three, and that the doctrinal lever of the volume tilts.

The Microsoft-Constellation agreement for the restart of Three Mile Island Unit 1 (Crane Clean Energy Center) was signed in September 2024 for 20 years, at 835 MW, with an initial commissioning target of 2028 brought forward to 2027 after the 18 November 2025 closing of a $1 billion loan from the Department of Energy to Constellation. The Amazon-Talen agreement on Susquehanna was expanded on 11 June 2025 to 1,920 MW through 2042, options to extend included, approximately $18 billion in contracted revenues, with a transition to a "front-of-the-meter" arrangement scheduled for spring 2026. The Meta-Constellation agreement on Clinton, signed 3 June 2025, covers 1,121 MW for 20 years from June 2027 on a site initially programmed for decommissioning, with a 30 MW capacity uprate. The Google-Kairos Power agreement, structured with the TVA, covers 500 MW of SMR, first site targeted for 2030. On the French side, EDF's EPR2 program, capped at €72.8 billion (2020 value) for its first six units, validated by the Board of Directors on 18 December 2025, targets a final investment decision by end of 2026, first concrete at Penly in March 2029, commissioning in 2038, and a 12 to 18 month cadence thereafter. The Nuward subsidiary, whose basic design for a 400 MWe SMR is expected mid-2026, explicitly targets energy-intensive industries and datacenters.

A frontier model generation lasts 12 to 18 months. A nuclear Power Purchase Agreement (PPA) commits 20 years. An EPR2 program commits the period 2029 to 2050 and beyond. *The deployer signs an energy horizon that outlives every model architecture it will ever train on it*. Energy sovereignty is not a decision about the current generation; it is a decision about generations N+1 through N+15, taken by a CTO who knows, at signature, neither the architecture nor the load profiles of the models to come.

It is here that the doctrinal point of the volume must be stated frontally, without softening. *A PPA is a decision longer than a model*. The consequence runs deeper than its formulation. From the moment AI becomes an industrial consumer of dispatchable power at the gigawatt scale, it ceases to be arbitrated primarily by the digital market. It enters the strategic arbitrage matrix of the energy state, where it competes on the same resource with decarbonized steel, industrial hydrogen, electrification of heavy transport, low-carbon district heating, semiconductor reindustrialization, and grid resilience requirements under NIS2. *AI governance, at this scale, ceases to be a software discipline. It becomes a variable of macro-industrial governance.*

This is a perimeter shift that none of the three previous volumes had to formulate. It does not invalidate the architectural doctrine; it subordinates it to an arbitrage that exceeds the deployer's perimeter.

CTO arbitrage matrix

Four constraints, four questions, four minimum proofs, four risks. The instrument is not a benchmark; it is an audit.

Constraint	CTO question	Minimum proof	Risk if absent
Allocated capacity	Does the site hold a connection agreement signed with the TSO, or merely a letter of intent?	RTE or equivalent national TSO agreement signed, indexed to a maximum drawable power, dated.	Commitment presented as guaranteed when it depends on a future SDDR; allocation reclassified in case of industrial prioritization.
Calendar	When is the source substation delivered, distinct from the datacenter opening date?	Energization schedule provided by the TSO, independent audit of critical milestones.	IT commissioning ahead of energization, operation on diesel gensets or captive gas.
Concordance	What share of the load is covered by long-term PPA (15 years and above), what share by wholesale market?	Declared supply contracts, weighted average duration, spot price exposure, time-stamped guarantees of origin.	Marginal inference cost not bounded over the contractual horizon, sustainability defect under CSRD.
Carbon intensity	What is the average and marginal carbon intensity of the localization grid, attested by an independent certifier and causally localizable?	EEA and Ember data at annual resolution, hourly GO certificates, documented Scope 2 GHG Protocol methodology, declared causal attribution of compute.	Environmental claim not defensible under applicable consumer and sustainability law.

This matrix separates what is proven from what is announced, and aligns the perimeter of the CTO decision with the perimeter of the regulatory commitments.

Three contemporary figures

The differential between American and European architectures is not a moral defect; it is an asymmetry of arbitrage perimeter.

1. **United States, internalization of generation.** Stargate announced in January 2025 at the White House, $500 billion of commitment, 10 GW of datacenter capacity targeted by 2029, of which 7 GW already allocated by September 2025. Anthropic announced a complementary $50 billion infrastructure investment in November 2025. Five major technology groups committed about $400 billion of capex in 2025, with announced 75% growth for 2026. The global pipeline of conditional SMR offtake has gone from 25 GW at end-2024 to 45 GW in May 2026 according to the IEA. When the grid cannot deliver fast enough, hyperscalers develop captive gas: Global Energy Monitor counts about 38 GW of gas capacity in development in the United States, of which roughly a quarter is dedicated to datacenters. *The AI load installs itself next to the power plant, or the power plant restarts for the load*. The American deployer's arbitrage perimeter includes the electrical generation decision itself.
2. **Europe, regulated composition.** Franco-Emirati AI campus 1.4 GW pending instruction; AWS European Sovereign Cloud generally available since 15 January 2026 in Brandenburg, €7.8 billion investment, operated under German law and scrutinized by the BSI; EPR2 program cadenced over 2029 to 2050; RTE offer of 35 ready-to-build sites. European investment is not lower in commitment level; it is different in structure. The European deployer purchases supply from a historical operator under regulatory and political oversight; it does not buy a power plant, and it does not trigger, by its decision alone, the construction of a new generation unit. It operates its PUE, negotiates its supply contract, optimizes its cooling, chooses its regions. It signs a calendar it did not write.
3. **Comparative reading.** The arbitrage perimeter is not the same. In the United States, the perimeter covers the AI load and the generation decision that powers it, sometimes on the same plot of land. In Europe, the perimeter covers the AI load and the contractual composition with a public actor whose planning exceeds the AI system in question. Both are choices of industrial architecture, and each produces its own constraints: private vertical integration on one side, regulated composition on the other. Volume 4 does not propose to replicate American vertical integration. It proposes to make European regulated composition legible, defensible, and instrumented. *A European CTO composing with RTE, EDF, and the Commission de régulation de l'énergie does not make a default choice for lack of a privatizable EPR. She operates within a perimeter where energy sovereignty is built as a composite good.*

European regulatory implications and causal non-localizability

Three regulatory corpora structure the energy defensibility of an AI system in a European environment. None of the three carries the same opposable force, and the confusion between them is itself a source of risk.

The GPAI Code of Practice, published by the AI Office on 10 July 2025, operationalizes the obligations of Article 53 of the AI Act. Article 53 obligations entered into application on 2 August 2025 for new models placed on the market from that date; models already on the market have until 2 August 2027 to come into compliance. The Code of Practice is voluntary, but adherence creates a rebuttable presumption of compliance. Its Transparency Section, articulated to Annex XI of the AI Act, requires every provider to complete a Model Documentation Form documenting architecture, training data, computational resources, and energy consumption of training. This documentation is retained for ten years, transmissible to the AI Office and downstream providers on motivated request. Article 53 sanctions reach €15 million or 3% of worldwide turnover.

Here appears the only useful point of irony in the volume, flagged as such. The final version of the Code of Practice introduces a clear-text exemption on energy disclosure. *Providers are exempted from disclosing the amount of energy used to train a model when they lack critical information from a compute or hardware provider*. The legislator acknowledges, in the very wording of the obligation, that the energy traceability of a model trained at a third party may be unavailable for lack of supplier cooperation. *An obligation whose own waiver is provided for by the legislator is not an obligation; it is a declaration of intent*.

This exemption is not a drafting curiosity. It reflects a structural problem the present volume names: *causal non-localizability of energy*. In a multi-region distributed hyperscaler chain, the marginal energy consumed by a given inference is not physically attributable with the precision regulators presume. Compute is dynamically routed across regions, GPU allocation is shared, the production mix varies hour by hour, guarantees of origin are fungible at annual scale, and provenance certificates are offset by purchases on secondary markets. The energy causal chain fractures across four actors: the model provider externalizes compute; the cloud provider externalizes electricity; the grid operator externalizes generation; the PPA externalizes the guarantee of origin. None of these actors holds, alone, a causal proof chain at hourly resolution and localized. And yet ESG claims, green-AI claims, CSRD obligations, and GPAI obligations rest implicitly on this attribution capacity.

This is the same pattern the previous volumes and earlier articles described for agentic orchestration: causal non-localizability produces a dilution of responsibility, which is treated by explicit signature or not at all. *The marginal energy of an inference is not localized; it is declared*. For a deployer in a regulated environment, the operative consequence is direct: defensible energy traceability must be obtained by bilateral contract with the infrastructure provider, by hourly guarantees of origin, by contractual signature of the provenance chain, or lost.

The CSRD inscribes training energy consumption within the perimeter of sustainability reporting for entities subject to its scope. On the field of consumer-facing claims, the regulatory landscape is in motion. The Green Claims Directive proposal, deposited in March 2023, was the subject of a withdrawal announcement by the European Commission on 20 June 2025 without formal withdrawal at the date of the present volume; the file is dormant. The effectively applicable instrument is the Empowering Consumers for the Green Transition (ECGT) Directive, adopted in March 2024, with national transposition due by 27 March 2026 and application from 27 September 2026. The ECGT prohibits in particular generic non-substantiated environmental claims, including carbon-neutrality claims based on offsets. As an order of magnitude, the average 2023 carbon intensity of the European electricity mix stands at 242 gCO2/kWh according to Ember, down 17% on the year and 40% on the decade according to the EEA, with considerable dispersion: Poland 662 gCO2/kWh, Czechia 450, Germany 371, Sweden and France among the lowest. *A model trained in Poland and a model trained in France are not the same regulatory object with respect to sustainability reporting*. The difference is defensible before a certifier, on condition of being documented and causally attributable.

NIS2, finally, does not "requalify" datacenters by their grid connection, contrary to a reading sometimes heard. Datacenter service providers are listed directly in Annex I of the directive (digital infrastructure): large enterprises in the sector (250 employees and above, or €50 million in turnover and above) are automatically *essential entities*, medium enterprises (50 employees or €10 million) are *important entities*, save for stricter national designation. The qualification is sectoral and size-linked, not connection-based. What plays at the energy interface lies in Article 21 of the directive: cyber risk management obligations explicitly include supply-chain security and critical dependency on other regulated sectors, of which energy also figures in Annex I. The operative consequence: a very large datacenter does not become essential because it consumes a lot; it already is essential by virtue of its size and sector. Its energy criticality translates into continuity obligations, dependency mapping, and incident notification covering its supply, opposable to the same national authorities (ANSSI in France, BSI in Germany) that supervise its cyber obligations.

The authentic limit: sovereignty under allocation

The most austere doctrinal point of the volume must be stated without softening.

An infrastructure with imperfect custody sovereignty (Volume 1) is compensated by encryption of data at rest and in transit, by key separation, by cryptographic registry of accesses. The defect is treated by instrument of proof. *Proof becomes the agent of sovereignty.*

A non-European silicon (Volume 2) is compensated by composite sovereignty: layer-by-layer audit, supplier redundancy, qualification of alternatives, industrial continuity plans. The defect is treated by structured diversification. *Diversification becomes the agent of sovereignty.*

A model with imperfect weight auditability (Volume 3) is compensated by a tripartite chain: inspectable and freezable weights, documented data, governed versions over time. The defect is treated by contractual and technical instrumentation. *Instrumentation becomes the agent of sovereignty.*

An absence of allocated energy is not compensated by an architectural instrument proper to the deployer. It is not encrypted, not diversified, not instrumented from the AI system side alone. It depends on an external system, the transmission grid, whose arbitrages are not piloted by the system architect. A site without its connection agreement on the planned opening date has no equivalent fallback. It can defer commissioning until the source substation is delivered, which misaligns every downstream commitment. It can switch to captive gas, which degrades carbon intensity and creates a defensibility problem under CSRD and the future ECGT regime. It can give up on the site, which invalidates the infrastructure investment. None of these outcomes preserves at once the industrial trajectory, the regulatory trajectory, and the environmental trajectory.

This is why the fourth doctrinal operator is neither proof, nor diversification, nor instrumentation. It is *allocation*. Energy sovereignty is a **sovereignty under allocation**, that is, a sovereignty whose central instrument is an attribution decision taken by a third party (the grid operator, in composition with the planning state), opposable to the deployer in long time. The cutting distinction at this final stage is between the three layers the architect instruments and the fourth she negotiates. *On the first three layers, sovereignty is what one proves. On the fourth, sovereignty is what one obtains.*

The predictable objection is that this asymmetry is temporary, and that it will dissipate as European generation strengthens. That is a hope, not an instrument. Doctrine does not run on hopes. Keystone pivot of this volume: *a PPA is a decision longer than a model*. Complementary physical pivot: *a GPU without an MW is not a GPU*. A CTO who has not instrumented these two distinctions, and who has not made explicit that the allocation decision partly escapes her, has not taken the measure of the system she pilots.

Conclusion: doctrine closed, audit open

The four-volume arc is now bound. European digital sovereignty is built on four layers, three of which compose by architectural instrument and one of which is requested from another.

The infrastructure layer is proven by a chain of cryptographic signatures. The fabric layer is diversified layer by layer along a three-verdict grid. The model layer is instrumented by an auditability-data-versions triad. The energy layer is allocated by a grid operator whose planning exceeds that of the digital system it powers. On the first three, the decision sits within the deployer's perimeter. On the fourth, it exceeds that perimeter and enters the strategic arbitrage matrix of the energy state, where AI competes with decarbonized steel, industrial hydrogen, electrification of heavy transport, district heating, and grid resilience requirements.

This is the civilizational consequence of Volume 4, and it will not disappear by architectural sophistication. *At the scales that structure the frontier, AI is no longer a subject of digital governance; it is a variable of macro-industrial governance.*

The next leap is no longer a fifth layer. The arc is complete. The next leap is the transformation of the composite doctrine into an executable instrument: audit protocol, signed arbitrage grid, qualification method opposable to an architecture review board or a notified body. Without this transformation, a doctrine, however accurate, runs the risk of every complete intellectual architecture: becoming coherent without becoming operative. The present series has tried to avoid that fate by anchoring each thesis in observable physical constraints, in dated regulatory regimes, in verifiable asymmetries, and in auditable instruments. The passage to executable audit is the test that will qualify the corpus, or fail to.

On energy sovereignty itself, the doctrine resolves to one sentence: *three layers are instrumented, the fourth is allocated*. And on the stack as a whole: *European digital sovereignty is composite, it is temporally asymmetric, and it is, in its fourth layer, under allocation*.