Policy Brief 18/2026

How can Europe address its pressing artificial-intelligence compute infrastructure shortfall?

Bertin Martens and Tillman Schenk

Bertin Martens (bertin.martens@bruegel.org) is a Senior Fellow at Bruegel

Tillman Schenk (tillman.schenk@bruegel.org) is a Research Analyst at Bruegel

Executive summary

ONLY TWO GIGAWATTS of artificial-intelligence compute capacity, equating to five percent of the global total, is located in the European Union. Planned projects will barely change the EU's global share. As local compute is a necessary condition for AI sovereignty, the EU should maximise domestic compute capacity, regardless of the origin of compute resources.

PRIVATE CAPITAL IS NOT THE CRUCIAL CONSTRAINT on Europe's AI compute build-out. Speed-to-operation and grid access are more significant limitations. An attempt is being made to address permitting through the European Commission's proposal for a Cloud and AI Development Act (CADA). Gigafactory subsidies could be redirected to competitive grants that put regions in competition with each other in terms of permitting times. Faster movers should get larger grants. Policy should also prioritise anticipatory grid investment and faster grid connections – grid access is left out of the scope of CADA.

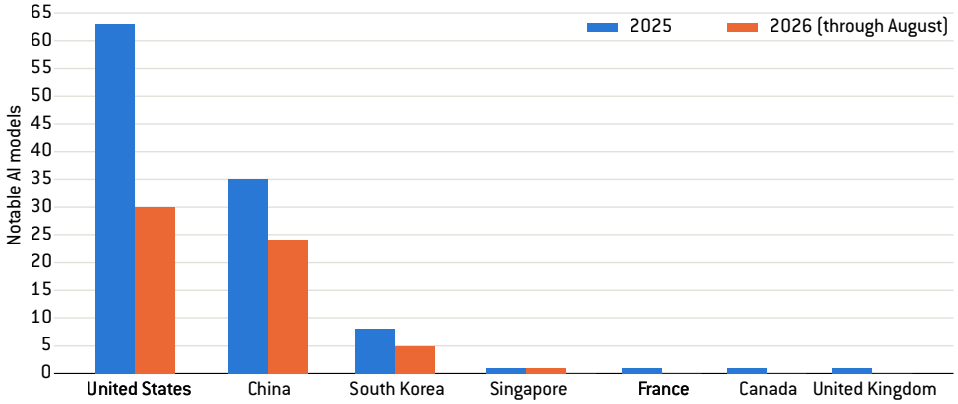
GLOBAL CHIP PRODUCTION CAPACITY CONSTRAINTS turn regional chip allocations into a near zero-sum game. The EU could incentivise United States hyperscalers to install larger shares of their AI chip supply in EU data centres. To this end, it should put sovereignty conditions on the AI services outputs side, for example through voluntary 'processed in the EU' labels on services sold to EU customers. Disclosure by hyperscalers of AI chips installed in the EU and AI workloads computed in the EU could also contribute. However, overly stringent local-content requirements, such as some of the proposed sovereignty criteria in CADA, risk raising costs and leading to a negative supply-side response.

Recommended citation
Martens, B. and T. Schenk (2026) 'How can Europe address its pressing artificial-intelligence compute infrastructure shortfall?', *Policy Brief* 18/2026, Bruegel, available at https://doi.org/10.64153/SHIA4474

1 The EU is running far behind the US and China in AI compute infrastructure

The European Union is running far behind the United States and China across the entire vertical artificial-intelligence stack, from chips and compute infrastructure[1] to models and applications. In the model layer, the EU accounts for barely any of the more than 100 notable AI models, and has just one significant model developer: Mistral. Global AI chip production is dominated by US firms – primarily NVIDIA, with Google in second place and a few other US players following – though China's Huawei is making inroads.

Figure 1: Number of notable AI models by country, 2025



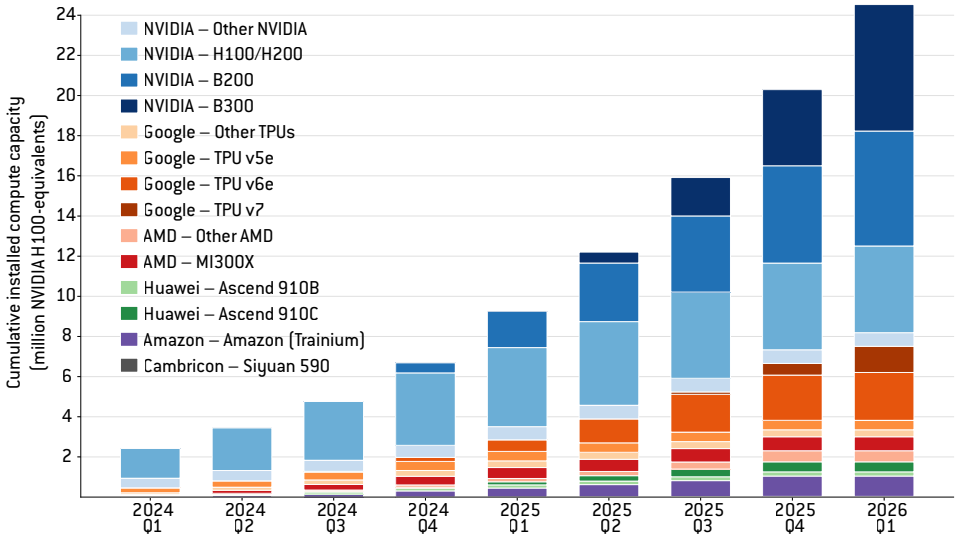
Source: Bruegel based on Epoch AI. Note: China excludes Hong Kong. A model is attributed to a country if at least one author is affiliated with an institution based there, so collaborations across countries count towards each one involved.

The EU is also assumed to lag in compute infrastructure, though before 2026 there was insufficient data to compare it with the US and China. There is very little transparency on AI infrastructure investment in China. In the US, by contrast, numerous sources track the exponential growth in hyperscale investment[2], driven largely by concerns about over-investment and its consequences for stock markets (Rungcharoenkitkul, 2026). It is less clear how much of that US AI compute is installed in US-owned infrastructure in the EU.

1 In the AI context, 'compute' refers to computing power or computational resources used to train AI models.

2 See for example Goldman Sachs, 'Global AI Investment Is Forecast to Exceed $1 Trillion in 2026', 7 August 2026, https://www.goldmansachs.com/insights/articles/global-investment-is-forecast-to-exceed-1-trillion-in-2026.

Figure 2: Global AI chip sales by cumulative compute capacity (H100e)*



Source: Bruegel based on Epoch AI. Note: * H100e = NVIDIA H100 processor compute capacity equivalents, a standard unit of measurement applied in AI chip comparisons. The NVIDIA H100 chip is an early-generation AI compute chip, dating back to 2022. NVIDIA cumulative totals include sales from 2022 onward. TPU cumulative totals include sales from 2023 onward. AMD, Amazon, Huawei and Cambricon have no published estimate for 2026 Q1. Their bars repeat their Q4 2025 total unchanged, not a projection of continued growth.

On the EU side, there were no comprehensive datasets on AI compute capacity until the publication in 2026 of the Europe2031.ai dataset of operational and planned EU AI compute centres (Juijn *et al*, 2026). This detailed list of 100 AI infrastructure investment projects of at least 10 megawatts[3] represents the first comprehensive picture of EU AI compute capacity. However, these figures exclude AI compute capacity in existing cloud data centres operated by US hyperscalers in the EU. There are no comprehensive datasets on this; only scattered press releases on AI chips installed in specific data centres.

According to this data (see Table 1), the EU in 2026 has about 2 gigawatts of operational AI compute capacity, accounting for nearly 5 percent of global installed capacity. The US has nearly 18 times as much (35 gigawatts, or 78 percent of global capacity), while China has more than twice as much (5 gigawatts, or 11 percent). While global capacity is projected to increase more than eight-fold by 2031, the relative distribution of these chips is not projected to change fundamentally. The EU share is forecast to increase only slightly to 5.6 percent, and China somewhat more to 15 percent of global capacity. Despite the striking figures reported regularly on US AI infrastructure investment, Juijn *et al* (2026) expect the US share of global AI compute to

3 The Europe 2031 study (Juijn *et al*, 2026) applies a minimum compute capacity of 10 MW to exclude a large number of much smaller AI compute infrastructures in private enterprises and universities, on which there is very little public information and that add little to the overall capacity. AI compute capacity is expressed in giga/megawatts, not in number of chips, which would distort the picture in favour of China, which uses less-energy-efficient chips, and older generations of energy-inefficient AI chips. 1 GW powers between 450,000 and 750,000 GPUs, depending on the efficiency. Forecasts do not take into account expected energy-efficiency gains in AI hardware. China has abundant and cheap electricity and does not face the same energy constraints on AI compute infrastructure as the US and EU.

decline from 78 percent to 68 percent, mainly because other countries are increasing their shares.

Table 1 shows that, with the current pipeline of planned AI compute infrastructure projects, including the EU's AI Gigafactories initiative (see section 3), the EU in 2031 will still be running far behind the US and China. It needs a quantum leap in AI investment to catch up. Taxpayer money cannot finance that leap; it will require far more private-sector finance.

Table 1: AI compute forecast 2025-2031 (in GW and percent)

	2025	2026	2027	2028	2029	2030	2031	2026	2031
US	17.27	35.28	61.25	104.37	146.56	208.20	255.13	78%	68%
Europe	1.43	2.07	5.17	11.76	12.83	16.80	20.89	5%	6%
China	2.53	4.95	9.84	19.11	28.98	45.00	55.95	11%	15%
Rest of the world	0.77	2.70	5.74	11.76	18.63	30.00	41.03	6%	11%
World	22.00	45.00	82.00	147.00	207.00	300.00	373.00	100%	100%

Source: Europe2031.ai (Juijn *et al*, 2026).

2 Does having AI compute capacity located in the EU matter?

In classic trade theory, economic efficiency would determine the optimal location of AI chips and compute capacity. Comparative advantage would allocate production geographically, increasing welfare for producers and users. Compared to the EU, the US is better endowed with the factors that go into AI services: skilled engineers, fabless chip designers and abundant domestic energy. Above all, US technology companies enjoy vast economies of scale: globally entrenched business models generate strong demand for AI services and the financial resources to cover the high fixed costs of building AI models and compute infrastructure.

China has similar advantages over the EU, though its technology companies are smaller and have less financial capacity than those in the US. On efficiency grounds, one could argue that EU firms should simply rent compute in the US or China and run their applications on infrastructure and models located elsewhere, assuming services trade flows without obstacles.

However, geopolitical tensions between the US and China, and between the US and its allies, have challenged the presumption of frictionless, politically neutral trade that the classic trade model rests on. The complex, interdependent, fragile global supply chains typical of digital products are vulnerable to geopolitical tensions and can be cut off and weaponised at very short notice by sovereign decisions. This uncertainty is compounded by the fact that digital products often operate on a permanent digital connection between producers and users. Cars, industrial machinery, data centres,

chip factories and closed AI models remain connected to their builders. Digital-service providers maintain the capacity to monitor performance after sales and to intervene at a distance for maintenance, updates and problem solving. That permanent connection and control over digital services can also be used to interfere with or stop user operations. Global connectivity is a conduit for benign and malign global interventions by all kinds of sovereign and private agents.

The rapid rise of AI as a general-purpose technology has lent fresh urgency to calls for more sovereignty, particularly in

This has injected considerable risk into the comparative-advantage model and prompted calls for 'sovereignty' and 'de-risking' the domestic economy[4] from malign interference by foreign actors or from *force majeure*, in particular in the digital domain where global connectivity constitutes a constant risk factor. US export controls on Anthropic's frontier models Fable 5 and Mythos 5, which forced the company to suspend model access,[5] are a powerful case in point. The location of supplies of critical products has become a relevant factor, further eroding the case for free trade and regional specialisation according to comparative advantages.

How much welfare are nations willing to forgo as an insurance premium against these risks? That question is particularly acute in AI. The rapid rise of AI as a general-purpose technology, with a strong US lead and China in second position, has lent fresh urgency to calls for more sovereignty, particularly in Europe. Former Italian prime minister Mario Draghi (2024) pointed out that AI compute capital will become an ever more important economic production factor – or token production capital, as distinct from human capital, according to Microsoft CEO Satya Nadella[6]. Lifting the EU's weak productivity growth will therefore require more access to AI compute, whatever its source. Because of geopolitical tensions, though, policymakers should rule out total reliance on American hyperscalers operating from US soil. Compute capacity inside the EU is a precondition for European AI sovereignty.

Other researchers argue that AI sovereignty depends on how much of the AI stack – and AI compute specifically – a country controls (Scharre, 2023; Sastry *et al*, 2024; Ding, 2024). Having a stake in the AI stack confers substantial economic and strategic value. Data-centre build-out generates short-term economic activity through construction spending and operational jobs, conveys technology spillovers and attracts AI researchers. Host regions can also tax data-centre revenue[7], which matters particularly if AI-driven automation shifts a substantial share of economic output from labour to capital. Beyond direct economic benefits, when an entity has sunk capital into a jurisdiction, it becomes easier to regulate and negotiate with it, and to leverage it (Poitiers and Schenk, 2026). For instance, model-access provisions can be secured

4 See for example Maria Demertzis, 'De-risking as an economic strategy', *Cyprus Mail*, 17 June 2023, https://cyprus-mail.com/2023/06/17/de-risking-as-an-economic-strategy/.

5 Mrinmay Dey, Jeffrey Dastin and Chris Thomas, "Anthropic disables top-tier AI models after US order limiting foreign access", *Reuters*, 15 June 2026, https://www.reuters.com/technology/us-blocks-foreign-access-anthropics-most-advanced-ai-models-axios-reports-2026-06-13/.

6 Satya Nadella, 'A frontier without an ecosystem is not stable', *X*, 14 June 2026, https://x.com/satyanadella/status/2066182223213293753. A token in the AI context is the smallest item of data, such as a word or part of a word, used for AI model training and inference.

7 Cecilia Kang, 'A County Got Rich From Data Centers. Some Question "At What Cost?"', *New York Times*, 21 August 2026, https://www.nytimes.com/2026/08/19/technology/data-centers-backlash-loudoun-virginia.html.

through contracts, and regulators can enforce local data-residency rules on model providers. Once Europe hosts a significant share of global compute, it also becomes much harder for the US to substantially restrict model access, even if the hyperscalers are American (Phillips-Robins *et al*, 2026).

3 The EU's AI supply-side approach

The EU policy response to its lack of compute involves multiple initiatives, including the AI Gigafactories (AIGF) initiative[8] and the proposed Cloud and AI Development Act (CADA; European Commission, 2026a). The AIGF would use taxpayer money to finance compute infrastructure, while CADA would establish 'sovereignty' criteria for public procurement and streamline permitting procedures.

The AIGF is supported by the InvestAI facility and aims to mobilise €20 billion of EU funds to co-finance €200 billion of private investment, to build up to five massive compute facilities, or 'AI gigafactories', dedicated to training frontier AI models[9]. In parallel, it supports 19 smaller AI factories and 13 'antennas', which do not have their own compute infrastructure but are linked to scientific supercomputers in existing European High Performance Computing (EuroHPC) centres[10], to be optimised for AI use with €2.6 billion in subsidies. By mid-2025, the European Commission had received 76 expressions of interest, involving private and public entities, for setting up AI gigafactories in 16 EU countries across 60 different sites[11].

The Europe 2031 list of all operational and planned AI compute infrastructure in the EU for the period 2025-2031 (Juijn *et al*, 2026) shows that 76 out of 101 projects are fully financed by private firms, accounting for 84 percent of all planned compute capacity. Only nine are publicly owned, and 16 have mixed private/public ownership, including the planned gigafactories. Ownership of these 101 projects is mostly European, though a few foreign private investors and sovereign wealth funds are involved in joint ventures with European companies.

The near-complete absence of US Big Tech firms from the list of 101 AI infrastructure projects – aside from two Microsoft projects[12] – is striking. A few smaller US neo-cloud firms are included. This supports our assumption that considerable AI compute capacity may already exist, unrevealed, in US hyperscalers' data centres in the

8 European Commission, 'AI Factories', https://digital-strategy.ec.europa.eu/en/policies/ai-factories.

9 See European Commission press release of 11 February 2025, 'EU launches InvestAI initiative to mobilise €200 billion of investment in artificial intelligence', https://digital-strategy.ec.europa.eu/en/news/eu-launches-investai-initiative-mobilise-eu200-billion-investment-artificial-intelligence.

10 The European High Performance Computing Joint Undertaking (EuroHPC) predates the AI Factories initiative; see https://www.eurohpc-ju.europa.eu/index_en.

11 See European Commission news of 30 June 2025, 'Overwhelming response as 76 respondents express interest in the European AI Gigafactories initiative', https://digital-strategy.ec.europa.eu/en/news/overwhelming-response-76-respondents-express-interest-european-ai-gigafactories-initiative.

12 The Europe 2031 list (Juijn *et al*, 2026) may underestimate US hyperscalers' investment in AI in Europe. They may build AI compute capacity in their existing cloud service centres; however, there is no transparency about these potential AI compute build-outs.

EU. For example, Google TPUs (tensor processing units)[13] represent about a quarter of global AI compute capacity (Figure 2). A significant part of that is likely to be installed in Google's EU data centres, for which there is no publicly available data.

At the same time, the contribution of the five planned gigafactories to the EU's total compute capacity is only 4 percent (about 750 megawatts, or 150 megawatts each) of the projected total of 21 gigawatts of AI compute capacity in the EU in 2031. Of the 101 AI compute centres planned in the EU by 2031, the gigafactories rank only 37th by power consumption. The largest centres are fully privately financed.

Several studies are critical of the AI gigafactories initiative. Kyosovska and Renda (2025) focused on poor location choices and lock-in to NVIDIA chips. Sieker and Hess (2025) argued that EU demand for AI services is too fragmented across many small users that cannot provide a stable source of scaled demand for large-scale compute capacity such as gigafactories. Avoiding underutilisation requires very flexible management that combines many sources of demand along the entire AI life cycle. Using taxpayer money to fund these subsidies is thus unwarranted, and the European Commission presents no economic justification for this investment.

The large number of privately funded data centres already in Europe suggests there is no lack of private interest in investing in AI infrastructure in the EU

The large number of privately funded data centres already in Europe suggests that there is no lack of private-sector interest in investing in AI infrastructure in the EU and no need for subsidies to stimulate private investment.

Another argument for subsidies is that the European Commission intends to allocate part of the AIGF compute capacity to small and medium-sized enterprises, start-ups, non-commercial public services and academic AI research[14]. But that too is unpersuasive. Compute capacity is not a public good in the economic sense, even when it is owned by public authorities[15]. Academic researchers and public services might as well rent compute capacity from the private sector and receive a subsidy to cover the rent. The only extra cost to taxpayers would be the provider's mark-up above marginal cost. As price competition among AI providers is intense, that mark-up is usually small, so renting could be cheaper for taxpayers.

Another often-invoked justification is that public-sector support will create 'sovereign' AI compute infrastructure – in particular data sovereignty – that does not depend on third-country suppliers and decision-makers. For example, the European Commission's 2026 Tech Sovereignty Package, of which the CADA proposal is a part, argues that AI gigafactories will ensure that "*industry, academia and public administrations are certain that strategic data and proprietary models remain under exclusive EU oversight to build the trust needed to migrate their most sensitive workloads*" (European Commission, 2026b, p 6). Granted that certain AI workloads could require oversight confined to the EU, the AI gigafactories will represent only 4 percent of EU compute capacity by 2031, and the sovereignty impact of that initiative

13 See Google, 'Introduction to Cloud TPU', https://docs.cloud.google.com/tpu/docs/intro-to-tpu.

14 See European Commission, 'AI Gigafactories', https://commission.europa.eu/topics/competitiveness/competitiveness-coordination-tool-projects/ai-gigafactories_en.

15 Public goods including clean air, defence and law and order are non-rival and non-excludable: nobody can be excluded from their use and benefits. Compute infrastructure is clearly excludable and rival, and therefore not a public good.

will thus be very limited. The remaining 96 percent, private and public-private AI compute investment, will contribute much more to local infrastructure in the EU, but still far from enough if the EU wants a share of global AI compute commensurate with its economic weight.

Locating AI infrastructure in the EU is a necessary but insufficient condition to preserve sovereignty. It makes it more difficult to cut the EU off from the infrastructure layer, though the AI model layer may still be cut off. This is where the proposed CADA comes in. CADA would establish 'sovereignty' criteria for compute infrastructure. It defines local-content requirements for a variety of AI hardware and software components (European Commission, 2026a). It sets out four cumulative levels of 'sovereignty assurance':[16]

- Level 1: The data-centre service provider and all subcontractors are legally established in the EU; all data is processed and stored in the EU only, and all outsourcing arrangements are traceable and do not compromise autonomy. Third-country service providers should be under no legal obligation to report vulnerabilities to their authorities.
- Level 2: Compute infrastructure, personnel and subcontractors are physically located in the EU. Personnel are EU citizens. Third-country subsidiaries are legally and technically separate from the EU parent company. Data is not used to train AI systems outside the EU. Third-country software components should not be able to disrupt the service.
- Level 3: No third-country access to customer data. Third-country involvement cannot restrain or restrict services. The audited provider should allow reasonable access to code.
- Level 4: Personnel with national security clearance. Auditors are established in the EU and not subject to control by third countries. Customer and audit data stay in the EU and are not used to train AI systems outside the EU. Open-source software components do not contain any remote features that could be used to tamper with the service.

Under CADA, public bodies would have to meet at least 'sovereignty assurance' level 1, and those with public-order relevance would have to procure levels 2-4. CADA conditions could be extended to private entities in critical sectors, such as transport, energy and health services.

These four levels of sovereignty criteria would progressively reduce the geographic scope of allowed locations for data centres and AI compute infrastructure, software development and personnel sourcing. By restricting the eligible supplier pool and adding auditing and certification obligations, CADA would raise providers' compliance costs, reduce competition and likely increase the price of compliant cloud services. Most US hyperscalers already offer 'sovereign' cloud services in the EU; these services

16 CADA requires adoption by the European Parliament and Council of the EU to take effect, and its provisions may be modified in the legislative process.

do not necessarily meet the four CADA criteria but go somewhat in that direction (Barczentewicz, 2026).

Governments operate under budget constraints. If the sovereignty criteria raise the unit price of CADA-compliant compute, public buyers will be able to afford less of it. The CADA criteria may thus slow down the AI build-out that EU investments seek to accelerate, thereby working against CADA's first proposed objective of tripling EU data-centre capacity by 2030. Public-sector buyers may be willing to pay a 'sovereignty premium' as insurance against foreign intervention. Whether private buyers would be willing to do the same remains to be seen.

There is thus no credible evidence of private-market failure for AI compute infrastructure in the EU that would justify substantial taxpayer subsidies for AI gigafactories. Moreover, the proposed gigafactories will not significantly increase the supply of AI compute capacity in the EU. The gigafactories tackle the EU AI compute problem from the financing side. However, capital is not the binding constraint, especially not for US hyperscalers with deep pockets. CADA local-content requirements for AI compute infrastructure may strengthen sovereignty beyond ownership and location, particularly for public services, but are also likely to raise prices and provoke a negative supply-side response.

4 What constrains the build-out of AI compute in the EU?

The current pipeline of planned EU AI compute infrastructure will not significantly increase the EU's share of global AI compute by 2031. Taxpayer-subsidised AI gigafactories will add very little to that infrastructure. The EU needs to accelerate its AI compute build-out significantly, but private investment, not subsidies, should finance it.

How can the EU attract more private investment for AI compute projects? Phillips-Robins *et al* (2026) concluded that the key factor determining how attractive a location is for data-centre investment is speed-to-operation: how long it takes from the first formal regulatory application to commencing operations. Based on a financial model that empirically estimates what factors determine the lifecycle value of a data centre[17], they found that speed-to-operation has a much more substantial effect on the rate of return on an AI infrastructure investment than other factors commonly assumed to matter, such as electricity prices, labour costs or taxes.

17 The Carnegie Endowment's International AI Data Center model (see Phillips-Robins *et al*, 2026) uses a discounted cashflow model of an illustrative 100MW greenfield hyperscaler datacentre (GB200 chips, NVL72 racks) over a 12-year operating life, run across ten countries. Revenues are built bottom-up from on-demand GPU rental prices of $10/hour at 70 percent utilisation, assumed to be identical across locations. Capital and operating costs are varied by country for construction costs, tax treatment and time-to-power. The datacentre is assumed to be funded from hyperscaler free cashflow, and all cashflows are discounted at a real weighted average cost of capital of 6 percent. Delay costs are derived by varying assumed time from construction start to first operation and recomputing lifecycle value.

In their model, a one-year delay costs an illustrative 100-megawatt data centre more than 5.5 percent of its lifecycle value. By comparison, a doubling of energy prices would cost the same illustrative AI data centre only 4.5 percent in lifecycle value. Phillips-Robins *et al* (2026) further found that Europe severely lags behind the US in time delays[18]. In the US, it takes an average of 24 months from the point at which a data-centre project has secured all necessary regulatory permits and confirmed how it will secure power to the point at which it first goes online. In Germany, the same process takes an average of 42 months.

European industry voices echo this. A 2026 survey by the European Data Centre Association, an industry body for data-centre operators, argued that "*growth* [in data centres] *is increasingly constrained not by capital or customer appetite, but by energy availability, grid readiness, and permitting complexity*" (EUDCA, 2026, p 4). *"Regulatory compliance plays an increasingly central role in site selection, design choices and long-term operational strategy, adding complexity*" (EUDCA, 2026, p 10). This suggests that permitting and grid access are the main current bottlenecks for the AI-infrastructure build-out.

Permitting typically requires complex coordination between national, regional and local authorities and private-sector entities, whose interests often diverge. Coordination failures are common and are, in practice, a bigger obstacle to investment than any private-sector finance failure. For instance, a variety of authorities handle land-use permits, environmental permits, grid connections and on-site electricity generation, without coordination or streamlined procedures. The CADA proposal recognises this problem and calls for the designation of Data Centre Acceleration Zones (European Commission, 2026a). EU national governments would designate specific sites where relevant national, regional and local authorities handle the necessary permits upfront. The relevant authorities would then grant an 'aggregated baseline permit', ie a single permit covering all permits commonly needed. Data centre operators in these Acceleration Zones would also get access to a Single Information Point – a one-stop shop that assists a data-centre project with all necessary authorisations.

However, CADA as proposed lacks a mechanism to enforce the proposed coordination structures and lacks incentives for the authorities to improve them. That is unsurprising: the EU does not exercise legal authority over regional and local authorities, or the relevant private entities. The Commission could, however, replace gigafactory investment subsidies with incentives for accelerated collective action in permitting and electrical grid connections. It could allocate grants to locations with the fastest permitting processes, which would put Data Centre Acceleration Zones in competition with each other. For example, subsidies could be scaled to how quickly a site completes permitting, with first movers getting the highest subsidies.

These subsidies could also compensate local opponents (NIMBYs) for any welfare losses caused by hosting AI compute infrastructure. However, this may not actually

18 The authors acknowledge high uncertainty in their findings from the underlying financial modelling. They include a sensitivity table showing that if assumptions are off by a few percentage points it may reverse the conclusions.

require EU subsidies. AI factories can be an important source of local tax revenue, which authorities can redistribute to local causes, for instance via tax reductions or cash transfers for people affected by nearby data centres.

Transparency and public communication about this allocation process – naming and shaming permitting laggards – could also motivate political authorities to implement efficient permitting procedures[19].

CADA as proposed would also fail to address the issue of electrical grid access[20]. In the regions where European data centres are clustered, the grid is already heavily congested. Cremona and Czyzak (2025) estimated that connecting a data centre to the grid takes seven to ten years on average in the markets where data centres are already concentrated – the so-called FLAP-D markets (Frankfurt, London, Amsterdam, Paris and Dublin) (EUDCA, 2026) – and that countries with less congested grids can expect twice the data-centre growth. Data-centre investors face similar constraints in the United States[21] (Gillet, 2026).

Could CADA successfully intervene to overcome regulatory coordination failure and accelerate permitting and grid connections sufficiently to enable the EU to achieve its proportionate share of AI compute? Accelerating the permitting of AI infrastructure projects currently in the planning pipeline by one year[22] would directly increase the EU's share of global AI compute capacity in 2031 from 5.6 percent to 6.7 percent. In addition, accelerated speed-to-operation would create financial incentives (Phillips-Robins *et al*, 2026) for US chip producers, hyperscalers and neo-cloud data centre investors to reallocate their chips from slower- to faster-permitting locations, which could further increase the EU's share of global compute.

In sum, the more binding constraint on AI-infrastructure development in the EU is a collective coordination failure between the relevant authorities. EU policies and funds for AI gigafactories should be used to incentivise those authorities, and the entities involved, to overcome it.

19 Maximilian Henning, 'An inside look at the EU's push to build AI gigafactories', *Euractiv*, 17 June 2026, https://www.euractiv.com/news/an-inside-look-at-the-eus-push-to-build-ai-gigafactories/.

20 David Janků, 'The Tech Sovereignty Package: Europe's Chance to Build Strength', *Arq Foundation*, June 2026, https://www.arq.foundation/research/the-tech-sovereignty-package-europe-s-chance-to-build-strength.

21 Chris Gillet, 'Why American data centers can't plug in', *Works In Progress*, 22 June 2026, https://worksinprogress.co/issue/why-american-data-centers-cant-plug-in/.

22 The second specific objective of CADA is to reduce data-centre permitting time to 18 months by 2030, compared to the current time of several years.

5 Global AI compute capacity

Even if Europe managed to accelerate the permitting process, bringing more AI chips into the EU would still hit global chip production capacity constraints. That shifts the problem to global chip reallocation mechanisms between countries and regions.

For the EU to get a larger share of global chip production, it must work more closely with US chip producers and their largest customers

The Europe 2031 study (Juijn *et al*, 2026) assumes that global AI chip production capacity is constrained by advanced 2-4 nanometre wafer manufacturing capacity in Taiwan at TSMC (Taiwan Semiconductor Manufacturing Company Limited), which is constrained in turn by the number of extreme ultraviolet lithography machines that ASML in the Netherlands can produce.[23] Since it takes years to expand these production capacities, we can assume that they are roughly fixed in the short and medium term. This binding constraint makes competition between US-based fabless AI chip producers a near zero-sum game played at TSMC's factory gate: one wins what the other loses. For the EU to get a larger share of global chip production, it must work more closely with US chip producers and their largest customers, US data-centre hyperscalers.

Fixed wafer production capacity does not mean that the demand for tokens goes unmet. Instead, it means that scarcity shows up in prices rather than in unserved demand. Token prices are flexible and competition between AI providers keeps demand and supply broadly in balance: as compute becomes scarcer, the opportunity cost of chips and of the services they produce rises, and users substitute towards cheaper, less compute-intensive models. Emberson and Sevilla argued that the shift to agentic AI is pushing that substitution along faster than compute capacity is growing.[24]

These forecasts only consider chips produced in the West, by US fabless chip producers, with ASML lithography machines and in the world's most advanced chip foundry in Taiwan. China has several AI chip producers, including Huawei, Cambricon, Alibaba and Kunlun. Their global market shares are very small (Figure 2). They are also less advanced than NVIDIA, mainly as a result of US export restrictions. This blocks their access to Taiwan's TSMC and the most advanced ASML lithography machines. Moreover, Chinese chip production is mainly reserved for domestic use, to compensate for US restrictions on NVIDIA exports – although limited exports to Chinese-financed data centres in South-East Asia and the Middle East have been reported[25]. Geopolitical tensions and the West's wariness about using Chinese AI hardware and software make a substantial increase in exports to the rest of the world unlikely. In addition, the dominant position of NVIDIA's CUDA AI model software makes transitioning to CUDA-incompatible AI chips difficult (García-Herrero and Martens, 2026).

When chip demand exceeds supply, the question is how to allocate that supply

23 This assumption is based on an earlier compute forecast by Kokotajlo *et al* (2025).

24 Luke Emberson and Jaime Sevilla, 'Is a compute crunch coming?' *Epoch AI*, 25 May 2026, https://epoch.ai/gradient-updates/is-a-compute-crunch-coming.

25 See for example Kai Nicol-Schwarz, 'The U.S. banned Nvidia's best chips from going to China. Now it's trying to close a crucial loophole', *CNBC*, 19 August 2026, https://www.cnbc.com/2026/08/19/china-ai-nvidia-chips-us-export-controls.html.

efficiently. In an open market setting, a price-based market-clearing mechanism would settle that. But NVIDIA and others sell their chips at fixed prices. That leaves substantial surplus demand that must be settled through discretionary rationing criteria, though whatever allocation choice producers make does not affect their revenue. NVIDIA claims that much boils down to first-come-first-served, but there has been no transparency or public communication about the criteria that determine this allocation[26].

Competition in the AI chips market is increasing. Google, Amazon and AMD are gaining footholds in the market (Figure 2). They are all subject to TSMC capacity constraints and must wrest quarterly production quotas from TSMC. They must also comply with US export rules when selling chips outside the US.

Google, the second largest producer of AI chips after NVIDIA, has long reserved TPU production for internal use and rental to enterprise clients, but has started to sell chips to external partners[27], giving priority to large contracts with major AI providers such as Anthropic and Meta, and to large private investors in AI compute centres[28]. Amazon reportedly follows a similar approach with its Trainium AI chips and is expected to start selling to major AI labs such as Anthropic, and to large cloud providers[29].

But competition in the AI chips market is far from perfect: differences in the software needed to run the chips make it hard for users to switch from NVIDIA GPUs to Google TPUs. The AI hardware and software layers are strong complements and cannot be considered separately. NVIDIA's strong CUDA software market position is due to network effects that bring together different types of AI users around the same multi-sided hardware platform: AI model developers and business-application developers; enterprises and service providers. This generates a large user community centred around the incumbent CUDA ecosystem, pooling experiences, solutions and application libraries. New market entrants, such as Google and Amazon, find it hard to break into this. Workaround software solutions are available but may not be as efficient as CUDA applications.

If the EU wants to increase its share of the global AI compute market considerably, it must ensure that a larger share of global AI-chip production actually ends up in EU data centres. That implies playing a delicate zero-sum game with US chip producers to convince them to allocate more chips to EU compute infrastructure. Accelerated permitting and grid connections may play an important role in this, more so than EU subsidies for AI gigafactories.

26 Dwarkesh Patel, 'Jensen Huang – TPU competition, why we should sell chips to China, & Nvidia's supply chain moat', *Dwarkesh Podcast*, 15 April 2026, https://www.dwarkesh.com/p/jensen-huang.

27 Georgia Butler, 'Google to sell TPUs to a "select group of customers" for their data centres, increases capex forecast', *Data Centre Dynamics*, 30 April 2026, https://www.datacenterdynamics.com/en/news/google-to-sell-tpus-to-a-select-group-of-customers-for-their-data-centers/.

28 Gyana Swain, 'Google opens TPUs to enterprises beyond its own cloud via Blackstone JV', *NetworkWorld*, 19 May 2026, https://www.networkworld.com/article/4173075/google-opens-tpus-to-enterprises-beyond-its-own-cloud-via-blackstone-jv.html.

29 Digital Applied, 'Amazon May Sell Its AI Chips: The Nvidia Challenge', 21 June 2026, https://www.digitalapplied.com/blog/amazon-custom-ai-chips-nvidia-challenge-2026-analysis.

6 What should Europe do about its AI compute shortfall?

Europe should prioritise and accelerate its compute build-out. Compute is essential to capture a larger share of the AI value stack and to accelerate productivity growth. But in a world of increased geopolitical tensions in which access to frontier models is not guaranteed, compute capacity is also a precondition for AI sovereignty. In attracting investment onto European soil, the EU should not discriminate against non-EU AI hardware inputs for its AI infrastructure projects on the basis of origin or ownership. It should prioritise more investment in EU AI infrastructure and make it more attractive for US hyperscalers to build in Europe.

More stringent local-content criteria for AI should wait until the EU has achieved a sufficient build-out of AI infrastructure

The criteria proposed in CADA risk the reverse: they would make it more costly to build infrastructure in the EU, potentially slowing down the shift in AI compute load from the US to the EU and making the EU more dependent on US-based AI compute. This will leave the EU with less sovereignty, not more. Instead, local-content requirements should target AI service outputs – the AI tokens consumed in the EU – for example by requiring compute services for EU customers to be located in the EU. That would avoid a negative supply-side response. More stringent and far-reaching local-content criteria should wait until the EU has achieved a sufficient build-out of AI infrastructure.

CADA also contains proposed provisions that could extend criteria on compute services to private procurement of AI services in critical sectors such as energy, transport and health. However, this should be voluntary, for instance by creating a 'processed in the EU' label for AI services. As awareness of the geopolitical risks of processing outside the EU increases, considerable customer demand for this quality label might emerge, provided hyperscalers are transparent about what the label means in practice, to avoid 'greenwashing' or 'EU-washing' compute workloads that are (partially) processed outside the EU. Hyperscalers would have to provide externally auditable evidence to support such quality claims.

Another way to bring more AI infrastructure into the EU would be to ask US hyperscalers to be more transparent about the extent of AI compute capacity and chips already available in their European cloud services. If this revealed a mismatch between EU AI workloads and their EU infrastructure – with much still processed in the US – it could nudge hyperscalers towards shifting more of their compute investment to the EU, helped by local-content quality labels.

It is also important to recognise that the binding constraint on EU AI infrastructure build-out is slow speed-to-operation, not financing. This is partly because of a grid-access bottleneck, which CADA leaves untouched. It is also because of failures in regulatory coordination between national, regional and local authorities involved in permitting. CADA's idea of Data Centre Acceleration Zones is a step in the right direction. To give regions further incentives to establish themselves as such zones, grants or tax breaks should be allocated to the zones with the fastest permitting

processes, thereby increasing competition between them.

The necessary EU-level funding for these grants could be redirected from the Gigafactory funds. The AI gigafactories push incorrectly identifies capital as the main bottleneck to the data-centre build-out. However, there is no market failure in private finance for AI infrastructure investment. Beyond a small set of strategic workloads, therefore, subsidies for EU AI gigafactories are not justified.

In addition, the EU should incentivise NVIDIA, Google and Amazon to allocate larger shares of their chips to AI compute infrastructure located in the EU by accelerating permitting and connections for that infrastructure, as argued above. The three leading US hyperscalers in the EU (Amazon AWS, Microsoft Azure and Google Cloud) account for about 70 percent of the European cloud market. Their estimated revenues (in 2025) from that market amount to approximately €27 billion for AWS, €20 billion for Azure and €13 billion for Google Cloud, meaning the European cloud market contributes roughly 20 percent of their global revenues[30]. Putting the same share of their €760 billion global AI infrastructure spending[31] in the EU would give a €120 billion – or 25 gigawatts – boost to EU AI infrastructure, more than doubling the EU's share of global AI compute by 2031, compared to the default scenario in Table 1.

US hyperscalers have good economic reasons to limit their investment in the EU, including issues relating to permitting and grid connection, and higher electricity and land prices. On the other hand, there is considerable consumer demand in the EU. AI compute infrastructure needs to be close to consumers to reduce latency, particularly for AI inference.

30 According to Synergy Research Group, the three US cloud hyperscalers – Amazon, Microsoft and Google – together accounted for approximately 70 percent of the $60 billion European market in 2024 (rising to more than $70 billion in 2025). These European market shares are not exceptional and closely reflect the global market shares of the cloud hyperscalers (AWS 28 percent; Azure 21 percent, Google Cloud 14 percent; total 63 percent). European-headquartered providers (SAP, Deutsche Telekom, OVHcloud, Orange) share most of the remaining 30 percent of the European cloud market. None report revenue for the European market separately. However, applying global shares to the European big-three revenue pool of 70 percent of €73 billion in 2025, gives the following rough revenue estimates for cloud services: AWS, €27 billion; Azure, €20 billion; and Google Cloud, €13 billion, or a combined market size of €60 billion. These are approximations, for several reasons. First, European market shares may differ from global shares, and the Synergy figures track infrastructure only. In relation to the latter, Microsoft's total European cloud footprint is actually much larger once the software-as-a-service subscriptions of European businesses are included; these could add another $20 billion to $30 billion on top of the Azure infrastructure estimate. See Synergy Research Group news of 5 February 2026, 'GenAI Helps Drive Quarterly Cloud Revenues to $119 Billion as Growth Rate Jumped Yet Again in Q4', https://www.srgresearch.com/articles/genai-helps-drive-quarterly-cloud-revenues-to-119-billion-as-growth-rate-jumped-yet-again-in-q4 and of 24 July 2025, 'European Cloud Providers' Local Market Share Now Holds Steady at 15%', https://www.srgresearch.com/articles/european-cloud-providers-local-market-share-now-holds-steady-at-15.

31 Katharina Buchholz, 'Big Tech's AI Spending to Reach $760 Billion in 2026', *Statista*, 31 July 2026, https://www.statista.com/chart/35046/capital-expenditure-of-meta-alphabet-amazon-and-microsoft/.

References

Barczentewicz, M. (2026) 'Europe's Sovereignty Stack: CADA, Compute, and the Limits of Autarky', *ICLE Issue Brief* 2026-06-15, International Center for Law & Economics, available at https://laweconcenter.org/resources/europes-sovereignty-stack-cada-compute-and-the-limits-of-autarky/

Cremona, E. and P. Czyzak (2025) '*Grids for data centres: ambitious grid planning can win Europe's AI race*', Ember, available at https://ember-energy.org/latest-insights/grids-for-data-centres-ambitious-grid-planning-can-win-europes-ai-race/

Ding, J. (2024) Technology and the Rise of Great Powers, Princeton University Press, available at https://press.princeton.edu/books/paperback/9780691260341/technology-and-the-rise-of-great-powers

Draghi, M. (2024) *The Future of European Competitiveness*, European Commission, available at https://commission.europa.eu/topics/competitiveness/draghi-report_en

European Commission (2026a) 'Proposal for a Regulation establishing a framework of measures for strengthening Europe's cloud and AI ecosystem (Cloud and AI Development Act)', COM(2026) 502 final, available at https://ec.europa.eu/newsroom/dae/redirection/document/129111

European Commission (2026b) 'Communication on European Tech Sovereignty, accompanied by an EU Open Source Strategy', COM(2026) 503 final, available at https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52026DC0503

EUDCA (2026) *State of European Data Centres 2026 report*, European Data Centre Association, available at https://www.eudca.org/new-2026-state-of-european-data-centres

García-Herrero, A. and B. Martens (2026) 'Stack battles: the US-China artificial-intelligence rivalry is moving beyond chips alone', *Analysis* 18/2026, Bruegel, available at https://doi.org/10.64153/CCJW6603

Juijn, D., S. van Baarsen, J. Dada, M. Negele, L. Stelling, P. Fox… M. Bakker (2026) *Europe 2031: What getting AI wrong means for us*, Europe 2031, available at https://europe2031.ai/

Kokotajlo, D., S. Alexander, T. Larsen, E. Lifland and R. Dean (2025) 'AI 2027', AI Futures Project, available at https://ai-2027.com/

Kyosovska, N. and A. Renda (2025) 'EU plans for AI (giga)factories: sanctuaries of innovation, or cathedrals in the desert?' *CEPS In-Depth Analysis* 2025-12, Centre for European Policy Studies, available at https://cdn.ceps.eu/2025/11/251027-Sanctuaries-or-Cathedrals.pdf

Phillips-Robins, A., T. Tawil and S. Winter-Levy (2026) '*The compute coalition: how to build the future of AI in the free world*', Carnegie Endowment, available at https://carnegieendowment.org/research/2026/06/the-compute-coalition-how-to-build-the-future-of-ai-in-the-free-world

Poitiers, N. and T. Schenk (2026) 'Revamping Europe's chips strategy: indispensability, not self-sufficiency', *Analysis* 12/2026, Bruegel, available at https://doi.org/10.64153/DADR3364

Rungcharoenkitkul, P. (2026) 'The AI investment race', *BIS Working Papers* 1367, Bank for International Settlements, available at https://www.bis.org/publications/working-paper-1367-ai-investment-race.pdf

Sastry, G., L. Heim, H. Belfield, M. Anderljung, M. Brundage, J. Hazell … D. Coyle (2024) 'Computing power and the governance of artificial intelligence', mimeo, available at https://doi.org/10.48550/arXiv.2402.08797

Scharre, P. (2023) *Four Battlegrounds: Power in the Age of Artificial Intelligence*, W.W. Norton, available at https://wwnorton.com/books/9780393866865

Sieker, F. and J.C. Hess (2025) 'Built for purpose? Demand-led scenarios for Europe's AI gigafactories', *Policy Brief*, Bertelsmann Stiftung, available at https://www.bertelsmann-stiftung.de/en/publications/publication/did/built-for-purpose

Bruegel, Rue de la Charité 33,
B-1210 Brussels
(+32) 2 227 4210
info@bruegel.org
www.bruegel.org

Republishing and referencing policy

© Bruegel 2026. Bruegel publications can be freely republished and quoted according to the Creative Commons licence CC BY-ND 4.0. Please provide a full reference, clearly stating the relevant author(s) and including a prominent hyperlink to the original publication on Bruegel's website. You may do so in any reasonable manner, but not in any way that suggests the author(s) or Bruegel endorse you or your use. Any reproduction must be unaltered and in the original language. For any alteration (for example, translation), please contact us at press@bruegel.org. Publication of altered content (for example, translated content) is allowed only with Bruegel's explicit written approval. Bruegel takes no institutional standpoint. All views expressed are the researchers' own.