EU PLANS FOR AI (GIGA)FACTORIES: SANCTUARIES OF INNOVATION, OR CATHEDRALS IN THE DESERT?

Nicoleta Kyosovska and Andrea Renda

CEPS IN-DEPTH ANALYSIS

November, 2025-12

SUMMARY

To address Europe's competitiveness delay in the domain of artificial intelligence (AI), the European Commission is investing heavily in building AI factories and gigafactories across Europe. They promise to create dynamic ecosystems for frontier AI development, bringing together compute, data, and talent. While acknowledging the Commission's efforts in driving infrastructure deployment, we examine whether it can meet its goals by investigating the implications of the factory locations, the type of architecture that is being built, and the type of AI that is likely to be deployed.

This CEPS In-Depth Analysis paper leverages data on patents, scientific publications, start-up investment, AI vacancies and electricity prices to analyse the conditions in the selection of factory locations. It finds that they are being built mostly outside of AI 'hubs of excellence' and that they are not leveraging the most favourable energy conditions in Europe. It also discusses whether, rather than focusing on generative AI and emulating the US approach at smaller scale, Europe should consider launching moonshots on alternative AI solutions and re-committing to its values by coupling the factories with dedicated research tracks on AI trustworthiness. Finally, it investigates the implications for Europe's sovereignty of relying solely (or mostly) on one technology provider for the supply of key components, including most notably AI chips.

Andrea Renda is Director of Research and Head of the Governance, Regulation, Innovation and the Digital Economy (GRID) unit at CEPS. Nicoleta Kyosovska is a Research Assistant in the GRID unit at CEPS.

The authors are extremely grateful to their colleagues from the Data Science unit at CEPS who provided the data for this paper, and especially to Gaia Cavaglioni and Francisco Ríos Fierro for their help in creating the visualisations. Please note that this analysis excludes the factories announced in October 2025.

CEPS • Place du Congrès 1 • B-1000 Brussels • Tel: (32.2) 229.39.11

CONTENTS

FIGURES

TABLES

EXECUTIVE SUMMARY

The EU has a competitiveness delay in the domain of artificial intelligence (AI), which has been acknowledged by the European Commission in several recent documents. A comprehensive plan was launched, featuring ambitious initiatives such as the 'AI Continent' and the 'Apply AI' strategies, coupled with a recent initiative for providing resources for AI in science.

One of the domains where the EU has been lagging significantly is compute infrastructure – enter the Commission's initiative to build AI (giga)factories. Starting with last year's announcement of the construction of AI factories – large sites with up to 25 000 advanced chips each, the Commission promised to create dynamic AI ecosystems, bringing together compute, data, and talent. This year, it decided to double down on this initiative by investing in up to five gigafactories, very large sites with at least 100 000 chips. Meanwhile, private sector giants such as Nvidia have announced gigafactory projects in various parts of Europe.

In this paper, we investigate whether the emphasis on compute infrastructure and the locations chosen for AI factories are likely to boost European competitiveness over the coming years. While acknowledging the Commission's efforts in driving infrastructure deployment, we also find important sources of concern in the Commission's current strategy. This is not only related to factory locations but also to what type of AI is likely to be deployed, how it will be deployed and why.

We leverage data on patents, scientific publications, startup investment, AI vacancies and energy costs to infer the logic that guides the Commission in identifying where the (giga)factories should be located. Our analysis involves the thirteen factories announced before October 2025. We find that they are being built in locations far from where Europe's AI 'hubs of excellence' are; and that places where factories are (or will soon be) located are not cooperating with each other. We then test the assumption that factories (like what happens in the US) are located in areas with favourable energy costs; yet we find that only factories in Sweden and Finland will benefit from comparable energy prices to those in the US and China.

We also discuss whether, rather than focusing on generative AI and emulating the US approach at smaller scale, Europe should consider launching moonshots on alternative AI solutions and re-committing to its values by coupling the AI factories with dedicated research tracks on AI trustworthiness.

Finally, we investigate the implications of relying solely (or mostly) on Nvidia for the factories' supply of graphics processing units (GPUs). We recommend that the EU diversify sources of GPU supply, mandates or prioritises open-source architectures, and

invests in alternative approaches that do not rely on GPUs, as well as in securing the supply of essential components such as memory chips.

To conclude, this paper hopes to shed light on how to achieve the daunting and conflicting goals of building competitive and sovereign AI by clarifying the conditions in which the AI factories should be built. This will enable Europe to make the most of its large investment. More holistically, our analysis aims to contribute to the urgent need for Europe to both distil its geopolitical position and disentangle it from its duty to build AI that will fully benefit society, thus becoming a true AI leader.

INTRODUCTION: FROM 'MEGA' TO 'BABE'

Over the past decade, governments around the world have become aware of the key role general-purpose technologies play in the competitiveness, security and sovereignty of their countries. This is particularly the case for artificial intelligence (AI), a family of technologies that lie at the core of new developments in science, advances in industrial production and distribution, societal relations, government services, health, defence and many other fields. While these technologies combine extraordinary opportunities with equally extant risks, having access to cutting-edge AI solutions has become an imperative for all governments.

In a globalised economy with free-flowing trade and global value chains, access to AI alone could be seen as sufficient to thrive. Yet today, an unprecedented crisis of global trade is prompting countries to reconsider their interdependencies and strengthen their strategic autonomy. This entails looking beyond access, and towards securing the ability to design and deploy AI at home, thus avoiding excessive dependencies on single sources of non-domestic supply. Or to use a popular term, promoting 'sovereign AI'.

Still, producing competitive AI systems at home is much more easily said than done. It takes cheap energy, raw materials, connectivity solutions, cloud services and infrastructure, abundant high-quality data, suitable institutions, adoption-ready markets and users, and world-class talent.

For Europe, the challenge is daunting as the continent has been de facto running on US digital solutions and Chinese raw materials for decades, and is today utterly dependent on foreign technology giants. Europe also struggles with high energy prices, the inability to retain or attract talent, and the chronic insufficiency of its capital markets. All this is exacerbated by the ongoing breathtaking race for AI leadership between the US and China, which leaves Europe caught in a dilemma. Should it follow the two more powerful rivals, and invest huge amounts of money, energy, water and compute to compete in the same field? Or should it follow the beat of its own drum and find a sustainable, trustworthy way to produce and use AI for the benefit of society?

Solving this dilemma is not going to be straightforward. The economics and the complexity of the AI age make binary solutions preposterous. Proponents of a completely autochthonous EuroStack envision a theoretical scenario, a full set of European solutions, which clashes with the immediacy of Europe's needs and thus represents at best an agenda item for the medium term. Defenders of the status quo ignore the magnitude of the threat: a sudden unavailability of raw materials from China, low-orbit satellite connectivity from Starlink, advanced chips from Nvidia, or cloud solutions from big tech would cripple Europe. Nonetheless, the recent US–EU trade 'deal', concluded under

threat of a US withdrawal of support for the defence of Ukraine, saw the EU renewing its commitment to buying American over the coming years. Today, somehow ironically, the 'magnificent seven' have already repurposed their business agenda, presenting themselves as messiahs of sovereign AI, also in Europe.

Rather than advancing towards the EuroStack, the sovereign AI solutions proliferating on the market may be a disguised iteration of the same tech dependencies. They imply data localisation and promise regulatory compliance but essentially remain high-consumption systems running on Nvidia graphics processing units (GPUs) and Broadcom application-specific integrated circuits (ASICs), US cloud infrastructure, and advanced GenAI solutions from OpenAI, Anthropic or Google.

At a time when the EU is dramatically under pressure to loosen both its digital and green regulations, this scenario implies abandoning any dream of digital grandeur or a 'European way' for the immediate future. There is no 'Brussels effect' or MEGA (Make Europe Great Again) strategy in digital, let alone the prospect of a fully-fledged European technology stack. Rather, the best Europe can hope for is BABE, 'Buy American' while gradually working to 'Build European'. This way, Europe would at least stand a chance of the next generation of solutions in the technology stack including the availability of European alternatives. But as explained below, this plan requires appropriate stewardship and clear investment choices.

In this paper, we look at one of the hottest fronts of the current AI race – the AI factories – and more generally, the rush to invest in compute. Since 2024, the European Commission has accelerated the construction of these large sites, initially scaling up the existing high-performance computing (HPC) infrastructure. In February 2025, Commission President Ursula von der Leyen doubled down by announcing that the EU will invest in three to five gigafactories, very large sites with at least 100 000 advanced chips.

We investigate whether the launch of these initiatives is likely to boost Europe's competitiveness in AI and related domains, and in what conditions. This depends, among other things, on the 'what', 'where', 'when' and 'how' factories are built. We also explore the 'what now', i.e. what should happen to enable the EU to make the most of this investment in the coming years.

The remainder of this paper is structured as follows. Section 1 describes ongoing projects to build factories in Europe (the 'what'). Section 2 assesses whether the number and location of the (giga)factories are justified from an economic and technology standpoint. Section 3 first addresses the 'when' by placing the current technological architecture in the context of extremely fast-moving market developments. Against this backdrop, the

proposed investment in high-power GPUs at scale seems to be challenged by parallel, sometimes alternative developments. It then discusses the 'how', drawing early insights on the compatibility of this plan with the EuroStack. The final section concludes by charting a 'European way' to AI.

1. WHAT IS BEING BUILT? A LOOK AT PUBLIC AND PRIVATE AI FACTORY PROJECTS

AI factories are defined as 'dynamic ecosystems that bring together computing power, data, and talent to create cutting-edge AI models and applications'. In January 2024, the Commission launched its AI Innovation Package to bolster start-ups and SMEs by offering privileged access to supercomputers already federated in the EuroHPC Joint Undertaking. Until then, the mandate of the Joint Undertaking was mostly about 'pure' compute; with the change, its mission was extended to include AI infrastructure and services (i.e. supporting AI model development) under the same umbrella. As many as 13 AI factories have been identified, sparsely dispersed across the territory of the EU (see Table 1 below).

In 2025, faced with the escalation of global investment in AI and the USD 500 bn Stargate Project in the (increasingly hostile) US, von der Leyen took the floor at the Paris AI Action Summit to announce an even more ambitious 'InvestAI' plan. It promises to leverage EUR 200 bn of (mostly private) investment and build a top tier of infrastructure sites known as 'gigafactories'. Each gigafactory will be endowed with approximately 100 000 state-of-the-art AI chips, to support the training of frontier AI models. This initiative has since become embedded in a broader AI Continent Action Plan, which targets up to EUR 20 bn of public investment for large-scale infrastructure and the overall strengthening of Europe's AI capacity. More recently, the 'Apply AI strategy' has relaunched the EU's efforts in this domain by laying the foundations for horizontal measures and domain-specific initiatives to boost AI uptake in strategic sectors and in government.

These developments will provide Europe with as many as 17 AI (giga)factories through public as well as public-private funding. However, this is not the end of the story: several private projects, mostly linked to US giant Nvidia often in cooperation with European companies, have been launched over the past few months. Nvidia has recently announced a roadmap to develop 20 factories in Europe, of which 5 are gigafactories. Table 1 summarises the existing projects to build factories in Europe, with 22 ongoing projects, to which a further 15 should be added given Nvidia's roadmap till 2030. It shows the location (where known) of each ongoing project, host institution, prospective compute availability, purpose or domain of application envisaged for the site, and timeline for its deployment.

Table 1. Public and private AI factories under construction in Europe

Country	Name	Host / location	Estimated compute	Purpose/domain	Timeline	Key players
Finland (+CZ, DE, EE, NO, PL)	LUMI	CSC – IT Center for Science (Kajaani)	Leverages LUMI-G ≈11 912 AMD MI250X GPUs; Factory adds AI services	Large-scale training Applications in: Health & life sciences Tech & digital Manufacturing & engineering	Operational; AI factory services ramping through 2025–2026	CSC + Nordic/CEE partners
Germany	JUPITER	Forschungszentrum Jülich (NRW)	Backed by JUPITER exascale (Booster ≈6 000 nodes; 4× GH200 per node)	Large-scale training Applications in: Health & life sciences Environment & sustainability Education & culture Manufacturing & engineering Finance & business Public sector	JUPITER live; AI factory services ramping 2025–2026	FZJ/JSC, ParTec, Eviden, SiPearl (cluster module)
Germany	HammerHAI	HLRS Stuttgart (consortium)	New AI-optimised system (details TBA)		Announced Dec 2024; implementation 2025–2026	HLRS + German HPC centres
Spain (+PT, TR, RO)	BSC AI Factory (MN5 upgrade)	Barcelona Supercomputing Center (Barcelona)	MareNostrum 5 AI upgrade (accelerator partitions TBA)	Large-scale training Applications in: Health & life sciences Tech & digital Environment & sustainability Education & culture Finance & business Agriculture & food Public sector	Operational by late 2025 (phased upgrades)	BSC-CNS, with PT, TR, RO partners
Italy	IT4LIA	CINECA @ Bologna Tecnopolo	>20 000 GPUs across AI-optimised systems (from 2026)	Large-scale training Applications in: Health & life sciences Tech & digital	2025–2027 rollout	CINECA, ACN, Emilia-Romagna, INFN, ICSC, universities

				Environment & sustainability Education & culture Finance & business Agriculture & food Cybersecurity & dual use		
Luxembourg	MeluXina-AI	LuxProvide (Bissen)	≥2 100 GPU-AI accelerators	Large-scale training Applications in: Tech & digital Environment & sustainability Finance & business Cybersecurity & dual use Space & aerospace	2025–2026 services buildout	LuxProvide, Luxinnovation, Univ. of Luxembourg, LIST
Sweden	MIMER	Linköping University (NAISS) & RISE	New mid-range AI supercomputer (TBA)	Large-scale training Applications in: Health & life sciences Tech & digital Environment & sustainability Education & culture Manufacturing & engineering Finance & business Agriculture & food	2025–2026 deployment	NAISS, RISE
Greece	Pharos (with DAEDALUS)	GRNET; national supercomputer DAEDALUS	DAEDALUS: 89 PF total; AI share reserved (TBA)	Large-scale training Applications in: Health & life sciences Tech & digital Environment & sustainability Education & culture	Starts 2025; 36 months	GRNET, NCSR Demokritos, NTUA, Athena RC
France	AI Factory France (around Alice Recoque)	GENCI / CEA TGCC (Bruyères-le-Châtel)	Exascale-class (Alice Recoque) – vendor details TBA	Large-scale training Applications in: Health & life sciences Tech & digital	System by ~2026; AI factory services 2026–2027	GENCI, CEA/TGCC, Inria, CNRS, partners

				Environment & sustainability Education & culture Manufacturing & engineering Finance & business Agriculture & food Cybersecurity & dual use Space & aerospace		
Austria	AI:AT (AI Factory Austria)	Advanced Computing Austria (ACA) & AIT (Vienna/TU Wien)	New AI-optimised supercomputer (numbers TBA)	Large-scale training Applications in: Health & life sciences Education & culture Manufacturing & engineering Finance & business Agriculture & food Public sector	Services Q4 2025; golive of the new system in Jan. 2027	ACA, AIT, TU Wien, Univ. of Vienna, partners
Poland	PIAST AI factory	Poznań (national hub; site TBA)	New AI-optimised system (TBA)	Large-scale training Applications in: Health & life sciences Tech & digital Environment & sustainability Space & aerospace Public sector	Announced Mar 2025; 2026–2027 rollout	National partners (health, cyber, space, sustainability)
Slovenia	SLAIF (Slovenia AI factory)	National host (Slovenia)	New AI-optimised system (TBA)	Large-scale training Applications in: Health & life sciences Tech & digital Environment & sustainability Education & culture	Announced Mar 2025; rollout 2026	SLING consortium, national ministries

				Manufacturing & engineering Agriculture & food		
Bulgaria	BRAIN++ (Discoverer++)	Sofia Tech Park (Sofia)	Next-gen Discoverer++ (GPU-heavy; numbers TBA)	Large-scale training Applications in: Tech & digital Environment & sustainability Education & culture Manufacturing & engineering Space & aerospace	Selected Mar. 2025; deployment 2025–2027	INSAIT, Sofia Tech Park
Germany	Industrial AI Cloud (AI Factory)	Deutsche Telekom (DE data centres)	≈10 000 Nvidia Blackwell GPUs (phase 1)	Enterprise AI	Announced Jun 2025; build 2025–2026	Deutsche Telekom, Nvidia (+ ISV ecosystem)
Sweden	Swedish AI Factory (enterprise)	New joint company (Linköping focus)	Phase 1: 2× Nvidia DGX SuperPODs (GB300); GPU count undisclosed	Domain-specific models, including for drug discovery and defence Large-scale inference for enterprise	Announced May 2025; under formation 2025	Wallenberg Investments, AstraZeneca, Ericsson, Saab, SEB, Nvidia
France, Italy, Switzerland, Spain, Norway	Telco Sovereign AI Programs		TBA (DGX SuperPODs & edge fabrics per telco)	Regional enterprise AI Edge/5G integration		Orange, Fastweb, Swisscom, Telefónica, Telenor + Nvidia
UK	UK GPU clusters (context)		120 000 Blackwell GPUs by 2026 per announcements	Enterprise/research AI		Nscale, CoreWeave (+ Nvidia)
Pan-EU (multiple including NO, SE, ES)	CoreWeave EU GPU campuses		'Thousands' of Nvidia GPUs (H100/H200/GB200 mix; site split undisclosed)	Frontier model training Enterprise AI	Sites online/ramping up by end2025	CoreWeave, Nvidia, customers including Mistral AI
TBD (site selection ongoing)	TBD	TBD (depends on vendor)	≈100 000 nextgen AI chips	Frontier model training; moonshots; scientific discovery	Formal call expected late 2025; build 2026–2028 (indicative)	Public--private partnership (PPP) under InvestAI

						(≈ EUR 20 bn total for 4 sites)
TBD (site selection ongoing)	TBD	TBD (depends on vendor)	≈100 000 nextgen AI chips	Frontier model training; very large European models	Late 2025 call; 2026–2028 build (indicative)	PPP (InvestAI)
TBD (site selection ongoing)	TBD	TBD (depends on vendor)	≈100 000 nextgen AI chips	Frontier model training; mission-critical apps	Late 2025 call; 2026–2028 build (indicative)	PPP (InvestAI)
TBD (site selection ongoing)	TBD	TBD (depends on vendor)	≈100 000 nextgen AI chips	Frontier model training; open innovation	Late 2025 call; 2026–2028 build (indicative)	PPP (InvestAI)

Note: this table does not include the factories announced in October 2025.

Altogether, as recently estimated, the 13 EuroHPC AI factories account for approximately 57 000 high-end AI accelerators, which is far too little considering how the market is evolving. To get a sense of the magnitude of existing investment in the US, consider that Meta planned to deploy infrastructure equivalent to nearly 600 000 Nvidia H100 GPUs by the end of 2024. And that, as shown in Table 1, the most recent investment announcement by Nvidia in the UK, as part of the 'Tech Prosperity Deal', entails the production of 120 000 GPUs and an investment of GPB 11 bn.

Even the gigafactories announcement pales when compared with investment levels in the US: Mark Zuckerberg recently announced his plans to have 1.3 million GPUs operating by the end of 2025. This helps put the gigafactories plan, with total of approximately 400 000 advanced chips, into perspective. It explains why private investment in factories, to the extent that it is well coordinated with publicly accessible factories, may still be the most important factor for Europe's competitiveness in the coming years.

2. HUBS AND (GIGA)FACTORIES – WHERE'S THE LINK?

Defined as dynamic ecosystems, rather than mere compute sites, factories are expected to mobilise the research community around AI development and deployment. As such, one would expect them to be built where suitable conditions exist. This intuition was recently supported by the Apply AI strategy under which a 'Frontier AI Initiative' was launched to bring together Europe's leading industrial and academic actors, leveraging cutting-edge AI architectures and high-quality data, as well as the computing capacity offered by the AI factories and gigafactories.

In this section, we look at the ongoing investment in AI factories and gigafactories. We draw on data from our past research on AI hubs (Balland and Renda, 2023; CEPS/UNIDO, 2024) and on new data from AI World – CEPS' new platform on AI which primarily uses data on patents, scientific publications, and start-up investment.

2.1. ARE AI FACTORIES BEING BUILT IN THE RIGHT PLACES?

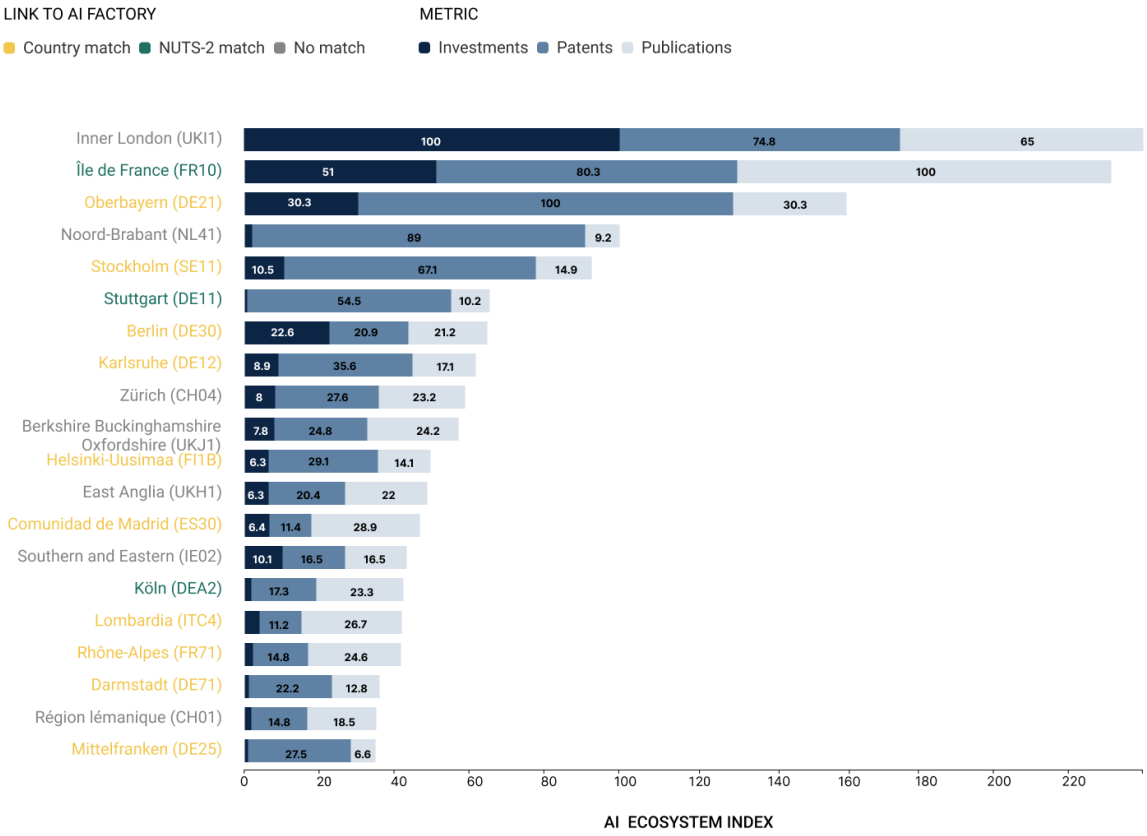
One of the most important criterions in the selection of sites suitable for the construction of AI factories is whether those places host a vibrant AI community and can thus be defined as 'hubs of excellence' in AI. Below, we approach this question by ranking European regions at the NUTS2 level according to a combination of the following metrics: the number of scientific publications in AI between 2021 and mid-2025 (source: OpenAlex); the number of patents in AI between 2021 and 2024 (source: EPO); and the magnitude of venture capital investment in AI start-ups between 2021 and mid-2025 (source: CrunchBase Pro). Each metric is scaled between 0 and 100 across the full period. The scaled metrics are then added together to form a single 'AI ecosystem index'. The AI index aims to proxy regional strength in AI R&I, and the depth of financial markets revolving around AI-based start-ups.

Figure 1 presents our results, showing the leading 20 AI hubs according to our bespoke AI index. The top hubs of AI excellence include Inner London (the city of London, UK), the Île-de-France (Paris, France), Oberbayern (the area of Munich, Germany), Noord-Brabant (the region of Eindhoven, the Netherlands), and Stockholm (Sweden). Five regions in the top 15 are in Germany. Besides those, the AI hubs are spread across 8 other countries (the UK, France, the Netherlands, Sweden, Switzerland, Finland, Spain, and Ireland), of which 6 are EU Member States.

The names of the regions are colour-coded to denote any link with a factory. Green indicates a region that hosts a factory; orange denotes regions in countries that host a factory (in another region); If the region name is black, this means that it is not linked to a factory in any of these ways.

The chart shows that there is only a small overlap between the leading AI hubs and the locations of the AI factories. These are the French AI Factory in the commune Bruyères-le-Châtel, in the region of Île-de-France where Paris is located, and the two German factories HammerHAI, at the University of Stuttgart, and JUPITER AI Factory, at the Jülich Supercomputing Centre in the town of Jülich, part of Köln.

Figure 1. Top 20 AI hubs based on scientific publications, patents and venture capital investments in AI start-ups, 2021 to mid-2025



Source: AI World

Notes: the top 20 AI hubs based on the number of scientific publications in AI for 2021 to mid-2025, number of patents in AI for 2021–2024 and amount (in USD) of venture capital investments in AI start-ups for 2021 to mid-2025. The AI ecosystem index is the sum of the normalised (0-100) metrics. The names of the regions (y-axis) are colour-coded to denote any link with a factory: green is a match with a factory; orange is a region in a country that hosts a factory; and black is none of these. See interactive version *here*.

Besides publications, patents, and investments, another very important factor in the selection of sites for AI factories is the availability of compute. To capture this element, Figure 2 adds an infrastructure dimension (y-axis) to the AI ecosystem index (x-axis), drawing on data on AI supercomputers from Epoch AI. This includes GPU clusters that were at least 1% of the size of the leading cluster at the time they first became operational. We therefore construct a 'compute index' by aggregating the total compute

capacity (measured in 16-bit TFLOPs) and the median energy efficiency (log of the 16-bit FLOPs per watt).[1] Both are standardised and combined as for the ecosystem index.

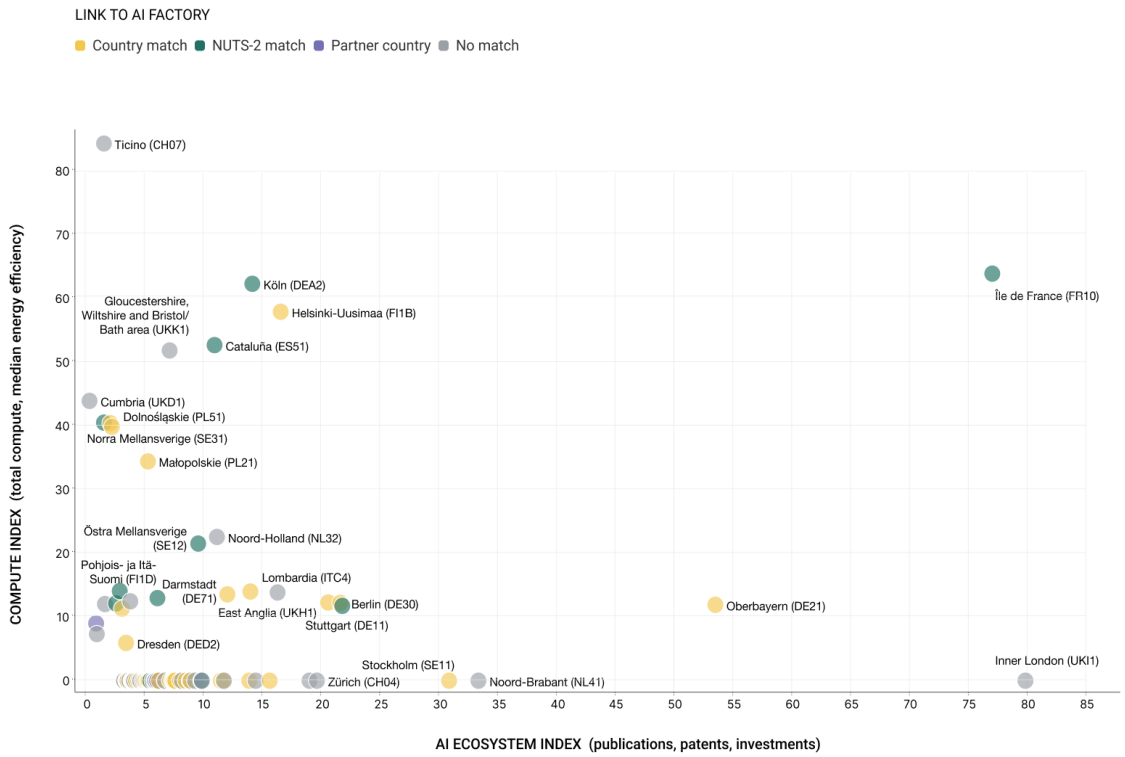
As with the colours of the region names in Figure 1, in Figure 2 the colours of the dots show whether a region is hosting an AI factory (green), whether a region is in a country that hosts a factory but the factory is in another region (orange), or whether the region is in a country that is a partner in an AI factory consortium, but does not host its own factory.

Figure 2 (see also interactive version here) includes all hubs, not only the leading 20 according to the AI index in Figure 1. This allows us to capture a richer set of findings. In particular, what emerges is that some regions, while not scoring high in terms of publications, patents and start-up investment, rank at the top of the compute index (where most regions appear at 0, so values above 10 can all be considered significant).

Based on this data, we show that leading locations in compute availability overlap slightly more with AI factory sites, compared with leadership in R&I. This is not a very strong correlation, but is visible in the graph as many green dots (i.e. regions where factories are located) fall in the top-left quadrant, where the score for compute is relatively higher than the score on the AI index. Apart from Île-de-France, Stuttgart and Köln, five more matches between regions and AI factory sites emerge, notably in Cataluña (with the Spanish AI factory in Barcelona); in Östra Mellansverige (with the Swedish AI factory in Linköping); in Wielkopolskie (with the Polish AI factory in Poznań); in Pohjois- ja Itä-Suomi (the Finnish AI factory in Kajaani); and in Emilia-Romagna (the Italian AI factory in Bologna).

[1] FLOPs – floating-point operations per second. TFLOPs – teraFLOPs - one trillion floating-point operations per second

Figure 2. AI ecosystem index vs the compute index: scores for leading regions and AI factory sites



Source: AI World for the AI ecosystem index; EpochAI for the compute index

Notes: the AI ecosystem index (publications, patents, and investments) is on the x-axis and the compute index (total compute capacity and median energy efficiency, source: EpochAI) is on the y-axis. The colour of the dots/regions denote any link with a factory: green is a match with a factory; orange is a region in a country that hosts a factory; purple is a region in a country that is part of an AI factory consortium; and black is none of these. See interactive version *here*.

Our data suggest that proximity to established and energy-efficient infrastructure is a key determinant in site selection, more than the availability of a vibrant AI ecosystem. This is confirmed by the recent investment decisions in the UK, where companies like Nvidia, CoreWeave, and Microsoft have partnered with Nscale to deploy AI factories mostly in northern England and Scotland. These regions possess some of the UK's richest renewable energy resources, particularly in wind and hydro power, and as such can offer cleaner, more sustainable power for energy-intensive AI operations, while also helping the UK meet its net-zero commitments. Many of these areas already have substantial grid infrastructure built to transmit renewable power southward, making them well positioned to support new, high-demand facilities. Land availability and cost also play a crucial role, leading the UK government to identify 'AI Growth Zones'.

For the EU, the most obvious pattern for the location of the factories is existing EuroHPC infrastructure. Out of the 13 factories, 9 will be built on sites with existing EuroHPC supercomputers. Among the participating countries, only Portugal and Czechia will not host factories. Most of these sites are in regions that do not qualify as excellent AI hubs according to our index but are located in countries that have hubs. On the other hand, some countries like Bulgaria, Greece, and Slovenia, while being designated hosts of AI factories, do not feature a region that leads in AI.

Our analysis shows that AI factories are distributed across multiple European countries, while leading European AI hubs are concentrated in just a few. This may be justified in order to ensure more correspondence between digital innovation hubs and AI factories, and also to give all Europeans a chance to access compute. At the same time, it risks the dispersion of resources in areas that are not well suited to the development of vibrant AI ecosystems.

Therefore, one aspect that will need to be clarified by the European Commission is whether the expectation is that AI factories themselves will host researchers and other stakeholders, giving life to an ecosystem of excellence, or whether factories will mainly be built in areas with compute availability and relatively low energy costs, with the R&I community able to access these resources largely remotely. This is important for Europe and is linked to the key question of whether investment should be concentrated in those few sites that combine high compute scores with AI index scores (e.g. Sweden, Finland, France, and Germany), or whether a more distributed network, perhaps with hubs and spokes (i.e. gigafactories and factories) is the preferred choice. Our research at CEPS shows that innovation tends to become more concentrated as the complexity of the technology increases.

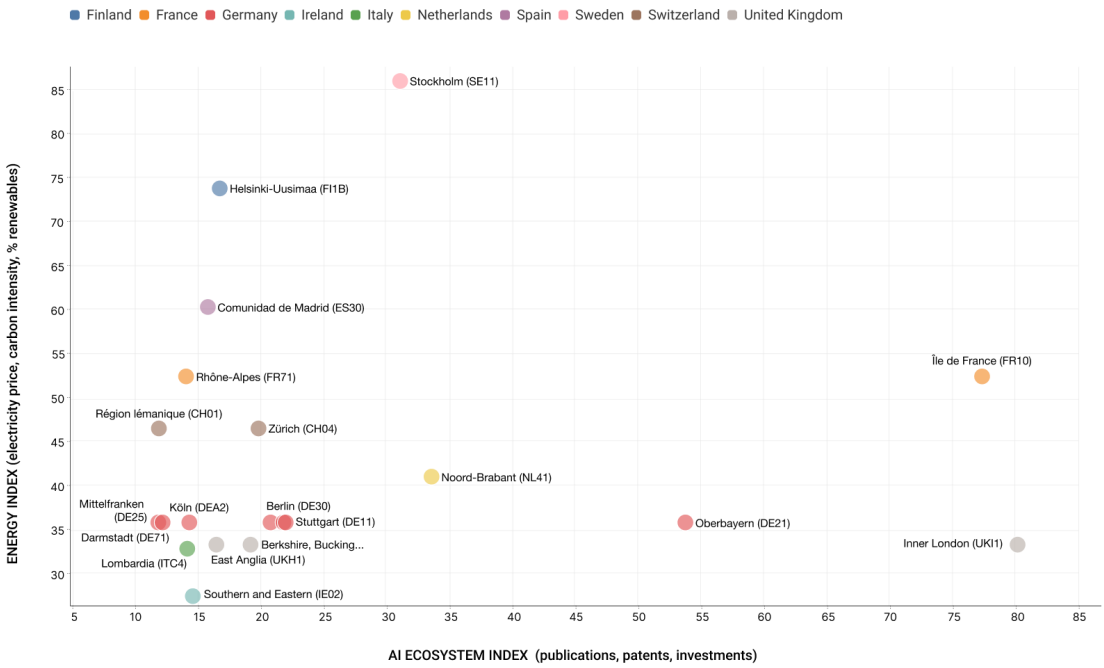
Moreover, even within countries with strong R&I activity, the AI factories largely match with sites that have highly efficient infrastructure, rather than with the leading AI hubs. The geographical dispersion of AI factories can be seen as a proxy for the deeper, systemic constraint of securing sufficient and efficient energy supply for AI data centres. To understand where the most suitable energy conditions in Europe are, we calculate an energy index at the national level. This combines three metrics:

- non-household electricity prices for 2024 (gov.uk for the UK; GlobalPetrolPrices for Switzerland; Eurostat for the rest);
- the share of renewables for electricity generation for 2023 (Electricity Map for the UK and Switzerland; Eurostat for the rest); and
- the carbon intensity of electricity generation for 2023 (Electricity Map for UK, Ireland, Norway, Switzerland; EEA for the rest).

Each metric is standardised to fall between 0 and 100. The reverse is taken (subtracted from 100) for electricity prices and carbon intensity. Finally, the three are summed to produce a single score.

European countries that score the highest according to the energy index are the Nordics – Norway, Sweden, and Finland. Figure 3 puts into perspective the top AI hubs (x-axis) and their national energy index (y-axis). We can see that the leading AI hubs are not located in countries with the best energy conditions, with the exception of the Swedish and Finnish hubs. While it is important to locate GPU clusters close to where data is served for inference – explaining why metropolitan areas typically host large concentrations of data centres (Data Centre Map) – distance is less critical for large-scale model training. This pattern is illustrated in the figure: many AI hubs lack significant local compute capacity, suggesting that they depend on remote machines for training activities.

Figure 3. AI ecosystem index vs the energy index of the 20 leading AI hubs (according to the AI index)

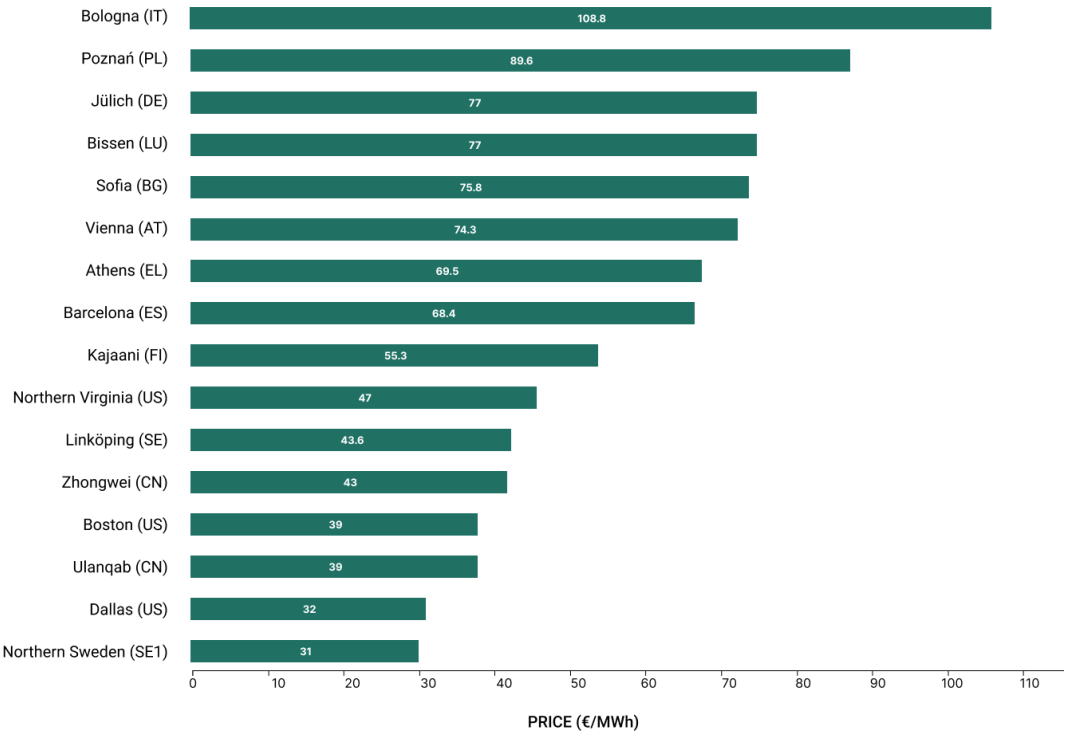


Source: AI World for the AI ecosystem index; Eurostat, ElectricityMap, gov.uk, GlobalPetrolPrices for the energy index (see the main text for details)

Notes: the AI ecosystem index (publications, patents and investments) is on the x-axis and the energy index (electricity prices, % renewables for electricity generation, and carbon intensity of electricity generation) is on the y-axis. Only the 20 leading hubs on the AI index are shown. The hubs are coloured according to their country.See interactive version *here*.

As appears from the data, the European Commission seems to be placing AI factories on sites with the most efficient infrastructure, rather than in AI 'hubs' located in countries that rank very high on AI research and innovation. For example, in Finland, the factory is located in Kajaani, one of the most eco-efficient areas in the world; the same can be said for Poznań in Poland. Placing the factories in these locations may be a sensible choice if one assumes that (i) failing to build compute infrastructure where energy costs are relatively affordable would be detrimental for AI's environmental impacts as well as competitiveness; and (ii) it is possible to build the infrastructure outside the excellence hubs, as researchers would be able to access compute infrastructure virtually. Figure 4 shows that only Swedish and (to some extent) Finnish AI factories are broadly comparable with select AI hubs in the US and China in terms of estimated energy prices (€/MWh).

Figure 4. Estimated energy prices for AI factory hubs (EuroHPC and others)



Source: Nord Pool for SE and FI; EPEX SPOT for DE, LU, AT and PL; OMIE for ES; GME for IT; HEnEx for EL; IBEX for BG; PJM IMM for Northern Virginia, US; ERCOT for Dallas, US; ISO-NE for Boston, US, NDRC for Zhongwei, China; NMG for Ulanqab, China

Notes: See interactive version *here*.

Current plans for AI factories in the US show a similar trend of separation between compute infrastructure and talent. Key US talent hubs include Silicon Valley, New York, Seattle, Boston, Austin, and (to some extent) areas around Washington D.C. and Virginia. Yet compute hubs are found in Northern Virginia ('Data Center Alley'), parts of Texas, Oregon, and Iowa, among others. In April 2025, the US Department of Energy identified 16 federal sites across the country (on federal land) that are considered promising candidates for developing AI / data centre infrastructure, due to existing energy or infrastructure assets. Other noteworthy projects include Microsoft's forthcoming data centre in Fairwater, Wisconsin for large-scale model training operations and Meta's investment in Holly Ridge, Louisiana. The USD 500 bn Stargate Project managed by OpenAI, Oracle and SoftBank involves the deployment of several new data centres in Texas, New Mexico, Ohio, and a Midwest site, with the aim of delivering up to 10 gigawatts of new AI computing capacity.

From an energy perspective, the emerging network of AI factories reflects a familiar imbalance between talent and infrastructure: it is generally inadvisable to locate AI factories within existing innovation ecosystems. This helps explain the limited overlap between leading AI research and innovation hubs and the locations of AI factories, at least in countries that already host such hubs. Should these nations still establish factories in alternative, energy-efficient areas? While this strategy may address the current energy demands of AI factories, it will fall short of meeting the far greater requirements of future gigafactories. Our analysis indicates that gigafactories should be more geographically concentrated – constructed within dedicated zones optimised for energy efficiency and supplied exclusively with additional renewable energy sources (Gröger et al., 2025).

2.2. WILL THE AI FACTORIES FOCUS ON THE RIGHT SECTORS?

Together with the announcement of the factories, the AI Action Plan published the sectors in which the factories will specialise (Table 2). By supporting the development of AI applications for specific sectors, the factories aim to boost competitiveness in strategic verticals. We looked at whether the designated domains correspond to priority areas for the regions hosting the factories, according to AI penetration and relative comparative advantage in R&I and start-up capital. High levels of AI penetration reveal sectors in which more efforts to adopt AI could bring large economic opportunities, given that their R&I activities already apply AI. The sectors' RCA reveals the likelihood of reaping the benefits: a high RCA indicates lower risk from applying AI, given a strong innovation ecosystem, while a low RCA indicates higher risk, since the capabilities are not yet present.

Table 2. Sector designations for the EuroHPC AI factories

Key Sectors	AT	BG	DE	EL	ES	FI	FR	IT	LU	PL	SE	SI
Health & Life Sciences	●		●	●	●	●	●	●		●	●	●
Technology & Digital		●		●	●	●	●	●	●	●	●	●
Environment & Sustainability		●	●	●	●		●	●	●	●	●	●
Education & Culture	●	●	●	●	●		●	●			●	●
Manufacturing & Engineering	●	●	●			●	●				●	●
Finance & Business	●		●		●		●	●	●		●	
Agriculture & Food	●				●		●	●			●	●
Cybersecurity & Dual use							●	●	●			
Space & Aerospace		●					●		●	●		
Public Sector	●		●		●					●		

Source: AI Continent Action Plan

The metrics are calculated as follows. We use publications for 2023–2024, patents for 2020–2024, and investments for 2021–2025, which are classified into GICS categories at the sub-industry level. We link the sub-industries to the sectors in which the factories will specialise (Table 2) and only show the results for those sectors that can be represented well. These are health and life sciences, technology and digital, manufacturing and engineering, finance and business, agriculture and food, and space and aerospace.

A region's RCA in a sector is the ratio between the sector's share of the region's total outputs and the sector's share of the EU's total outputs. We calculate the RCA in publications, patents and investments separately, normalise them to be between 0 and 100, and take their average to build a single RCA score. This score indicates the regional concentration of R&I and start-up capital in the sector relative to the EU. AI penetration in a sector is calculated as the share of publications/patents/investments in a sector that are linked to both this sector and AI. Similarly to the RCA score, we have taken the average of the normalised AI penetration rates across the three metrics to build a single AI penetration rate. It indicates the share of R&I activity in the sector where AI is being applied.

Figure 5 below includes a sub-plot per hosting region, showing the RCA (x-axis) and AI penetration (y-axis) of the six selected sectors: health and life sciences, technology and digital, manufacturing and engineering, finance and business, agriculture and food, and space and aerospace. The sectors are grouped into four quadrants, based on whether they are above or below the median for the region across these six sectors. We consider sectors in the top-right quadrant to have the highest strategic importance for the region, since they combine high AI penetration with high RCA, which indicates they are investment zones with high reward and low risk.

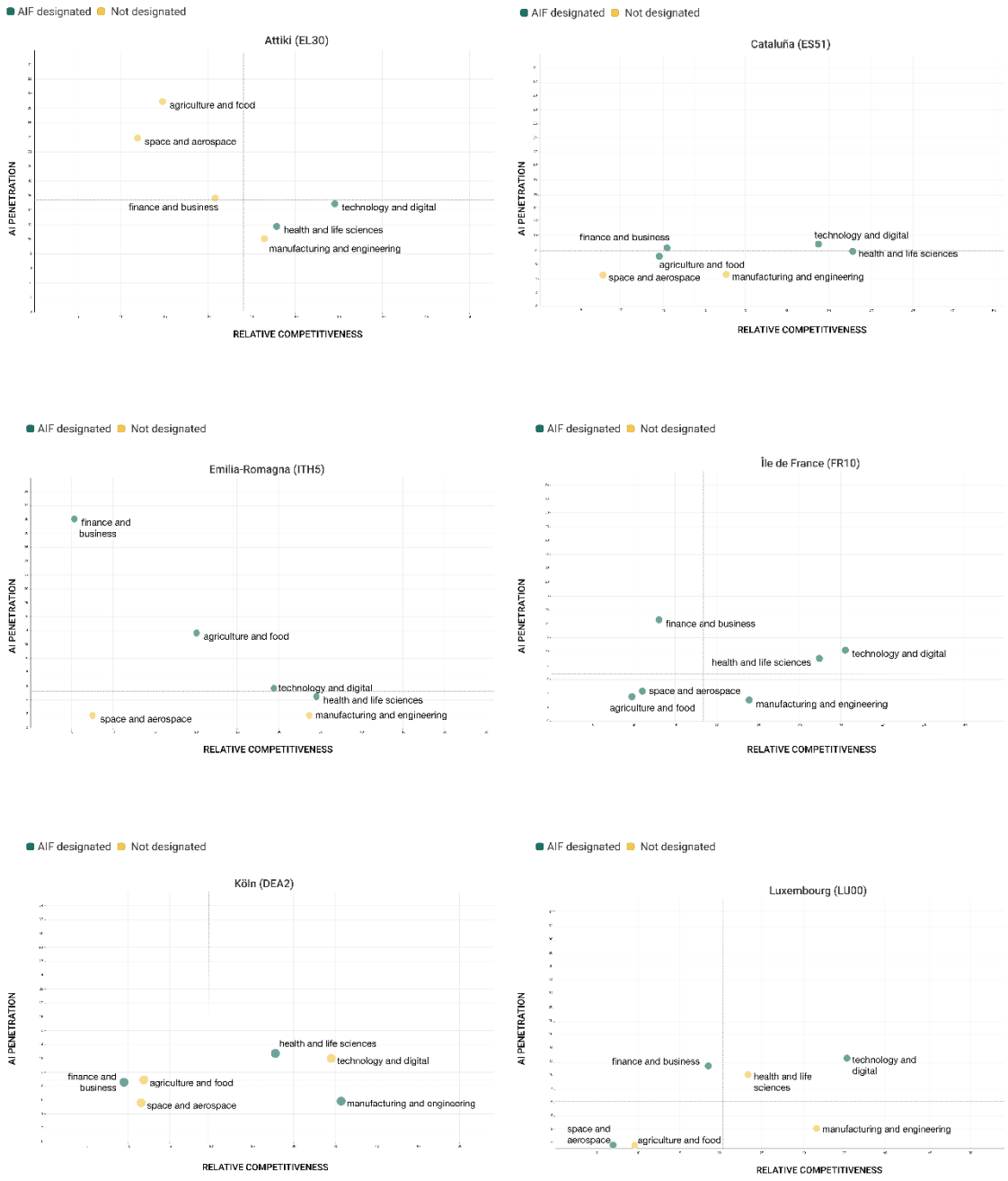
We overlap the results with the factory designations, finding that the data backs up some of the selections, such as for the Spanish and Finnish factories. The region of Cataluña shows high AI penetration rates in the designated finance, tech, agriculture, and health sectors which are all also above or close to the median. Out of the three designated specialisations for the factory in Kajaani, manufacturing and technology see high RCA and AI penetration close to the median, and health is well above the median on both measures.

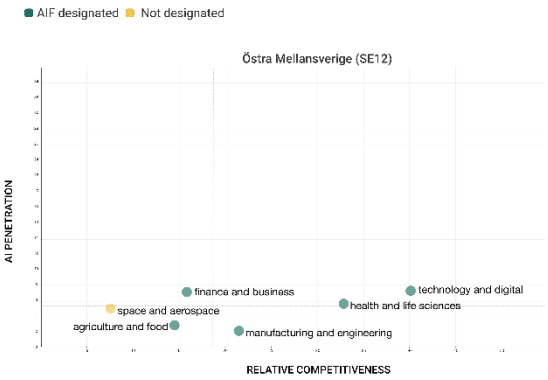
Other factories show a tendency to specialise in sectors in which their regions are either competitive (concentrated in the quadrants on the right) or have high AI penetration (concentrated in the top quadrants), but not both. This is the case for the Slovenian factory, with their sectors showing high competitiveness but varying AI penetration rates, agriculture being significantly low. By contrast, the sectors in the Italian hosting region score above the median AI penetration rate but agriculture and finance rank below the median RCA.

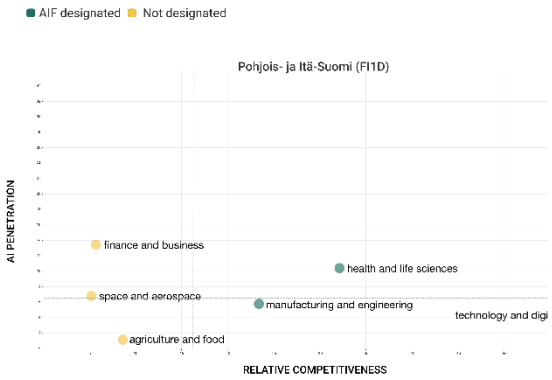
The following other patterns are notable. Some regions exhibit sectors with high potential that are not designated, for example, finance and business is in the top-right quadrant of the Greek factory. Conversely, the factory in the region of Cologne is designated finance and business, but is actually more competitive in technology and digital, which also has a higher AI penetration rate. Technology and digital is in the top-right quadrant for almost all regions, which is not surprising given the presence of compute infrastructure in these regions. Finally, some factories, such as the Swedish and the French ones, have many designations, which are not all strategically important according to our measures.

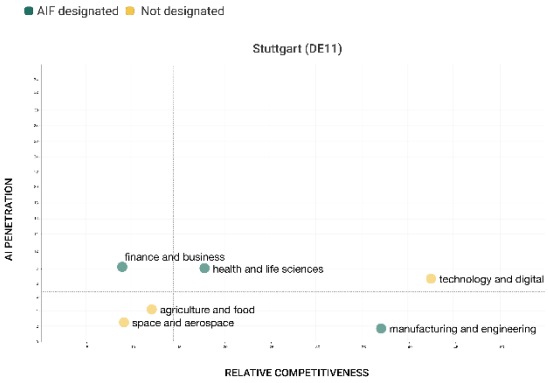
To conclude, the factories should specialise in sectors that align more strongly with current levels of AI penetration and comparative advantage, to maximise regional competitiveness.

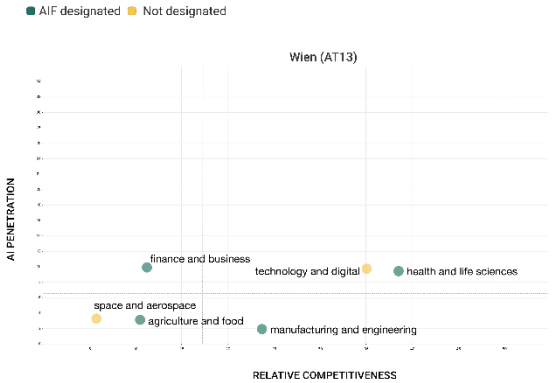
Figure 5. Difference in RCA vs AI penetration for six of the factory-designated sectors in each AI factory region

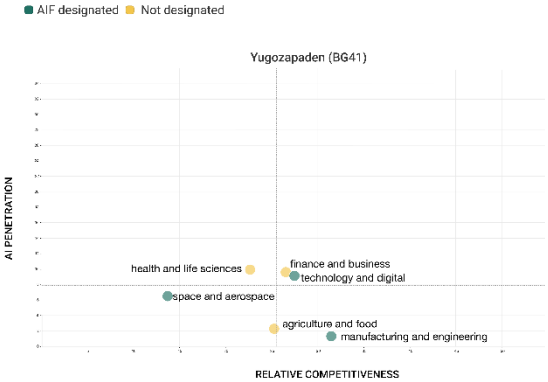


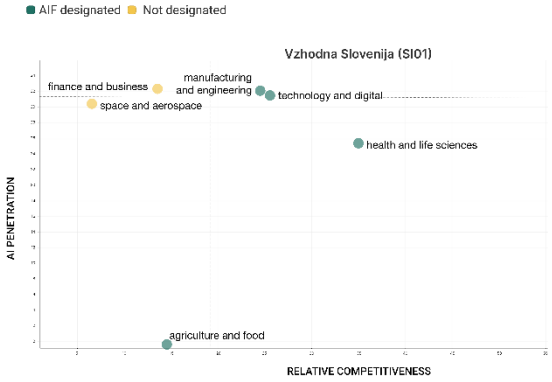

AIF designated
Not designated
Östra Mellansverige (SE12)
AI PENETRATION
RELATIVE COMPETITIVENESS
finance and business
technology and digital
space and aerospace
health and life sciences
agriculture and food
manufacturing and engineering

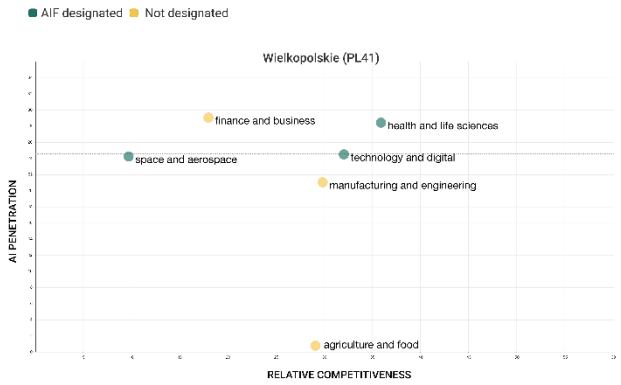

AIF designated
Not designated
Pohjois- ja Itä-Suomi (FI1D)
AI PENETRATION
RELATIVE COMPETITIVENESS
finance and business
health and life sciences
space and aerospace
manufacturing and engineering
technology and digi
agriculture and food


AIF designated
Not designated
Stuttgart (DE11)
AI PENETRATION
RELATIVE COMPETITIVENESS
finance and business
health and life sciences
technology and digital
agriculture and food
space and aerospace
manufacturing and engineering


AIF designated
Not designated
Wien (AT13)
AI PENETRATION
RELATIVE COMPETITIVENESS
finance and business
technology and digital
health and life sciences
space and aerospace
agriculture and food
manufacturing and engineering


AIF designated
Not designated
Yugozapaden (BG41)
AI PENETRATION
RELATIVE COMPETITIVENESS
health and life sciences
finance and business
technology and digital
space and aerospace
agriculture and food
manufacturing and engineering


AIF designated
Not designated
Vzhodna Slovenija (SI01)
AI PENETRATION
RELATIVE COMPETITIVENESS
finance and business
manufacturing
and engineering
technology and digital
space and aerospace
health and life sciences
agriculture and food



Source: CEPS

Notes: relative competitiveness on the x-axis is the relative competitive advantage (RCA) in the publications, patents and investments in the domain. It is calculated separately for each metric, then normalised to be between 0 and 100, and taken the average of. It indicates the regional concentration of R&I and start-up capital in the sector relative to the EU. AI penetration on the y-axis is the share of publications, patents, and investments in the domain and also in AI. It is calculated across the three metrics in the same way as the RCA. See interactive version *here*.

2.3. NETWORKS OF COLLABORATIONS BETWEEN HUBS AND FACTORIES

Beyond the energy and compute dimensions, the existence of R&I collaboration is also an important condition for the design of a network of AI ecosystems. Below, we use network analysis to understand how much the sites selected for the AI factories collaborate with each other, as well as with leading European hubs, compared with what could be expected based on their size and degree of connection – the number of direct connections they have in the network (Balland et al., 2025).

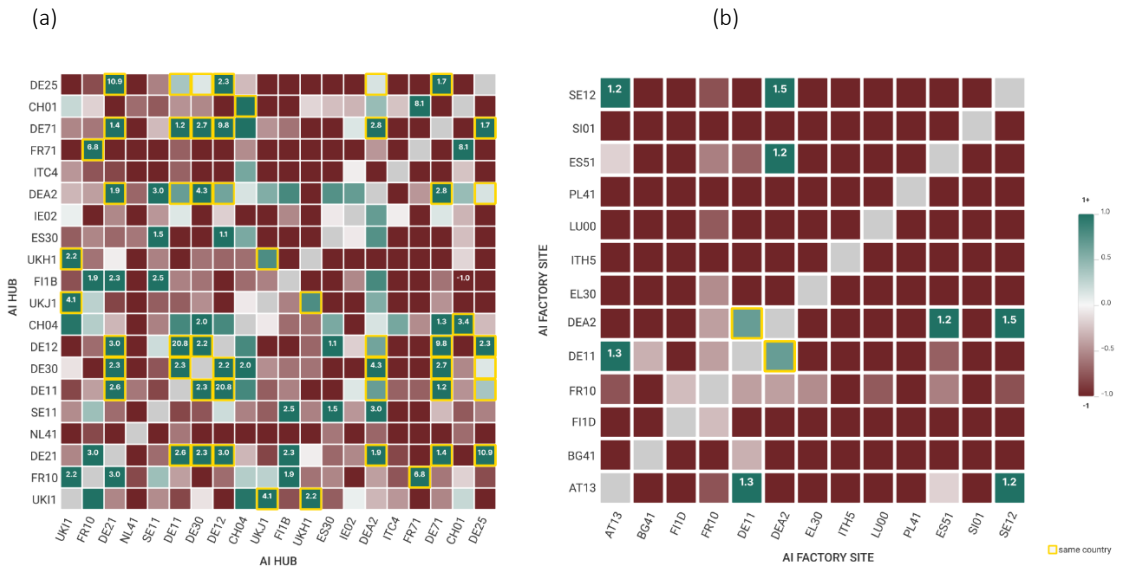
Figure 6(a) shows the extent to which actual levels of collaboration on patents deviate from the expected level among the leading 20 hubs in patents from 2021 to 2024. The metric shows how many patents were more or less co-created as a proportion of the expected number; for example, if the regions of London and Paris were partners in 20 patents, and the expected number was 10, the difference in RCA would be 1.[2] A value of -1 means there were no observed collaborations leading to patents; a value of 0 means expected and actual were the same. What we observe is a portrait of Europe's still rather fragmented single market: most of the hubs within the same countries tend to collaborate more than expected (denoted with a dot on the heatmap); otherwise, there

[2] The formula is n_actual/n_expected – 1 = 20/10 – 1 = 1.

is usually untapped potential for partnerships, e.g. between Stockholm (SE11) and Stuttgart (DE11).

Figure 6(b) shows the same measure, but this time for the network of regions where the AI factories will be located. Links between these regions are much less exploited, with most of them lacking any co-created patents (the difference in RCA being -1). The regions that lead in AI overall – Stuttgart (DE11), Köln (DEA2) and Île-de-France (FR10) (Figure 1, in rows 4 to 6) – are also engaged slightly more than the rest. Outside these, there is only a strong link between Wien (AT13) and Östra Mellansverige (SE12). The regions of the Italian, Polish, and Slovenian AI factories have not co-authored patents with any of the rest.

Figure 6. Difference in the RCA of actual and expected collaborations among the top 20 European regions for AI patents (a) and among AI factory regions (b)



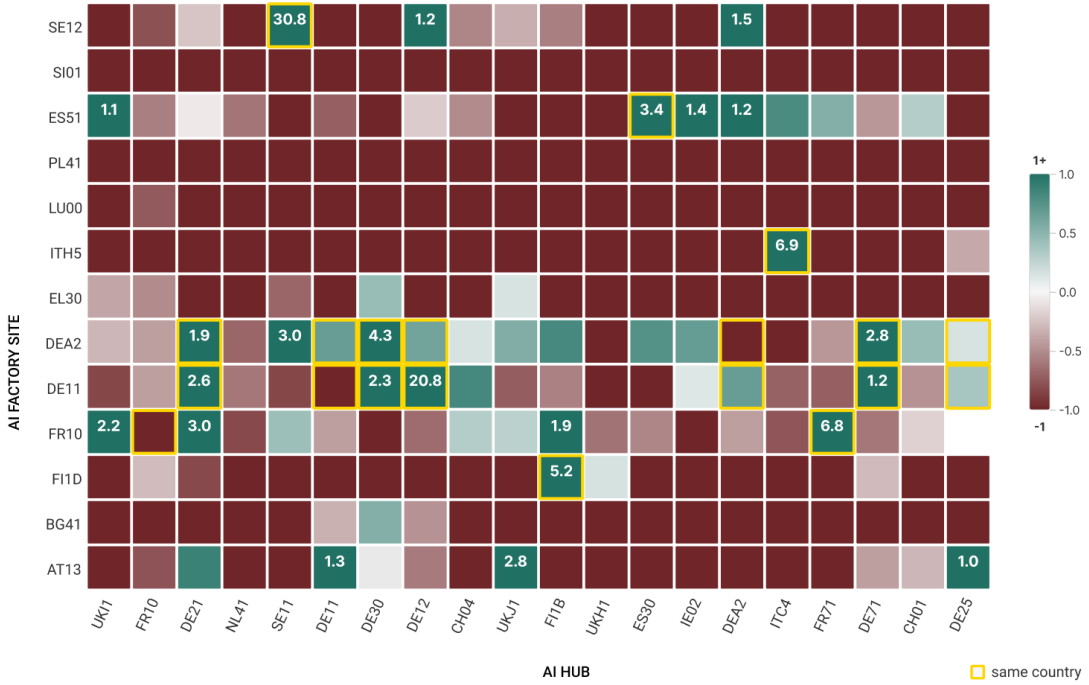
Source: CEPS

Notes: the metric shows how many patents were more or less co-created as a proportion of the expected number; for example, if the regions of London and Paris were partners in 20 patents, and the expected number was 10, the difference in RCA would be 20/10 – 1 = 1. A value of -1 means there were no observed collaborations; a value of 0 means expected and actual were the same. The positive values are in green, while the negative ones are in red. The lowest possible is -1, as collaborations cannot be negative. See interactive version for a) *here* and b) *here*.

Figure 7 below presents the pattern of activity between the AI factory sites and the leading AI hubs. The resulting picture shows a somewhat more developed network than the one related to AI factory sites (Figure 6(b)), but still displays significant untapped potential. Similar to the previous case, the regions that stand out with more activity also lead in AI overall (Figure 1): Stuttgart (DE11), Köln (DEA2) and Île-de-France (FR10).

Another site that stands out is Cataluña (ES51), which is strongly connected with several patent hubs, such as Southern and Eastern Ireland (IE02), Lombardia (ITC4), and Rhône-Alpes (FR71).

Figure 7. Difference in the RCA of actual and expected collaborations between AI factory regions and the top European hubs in patents

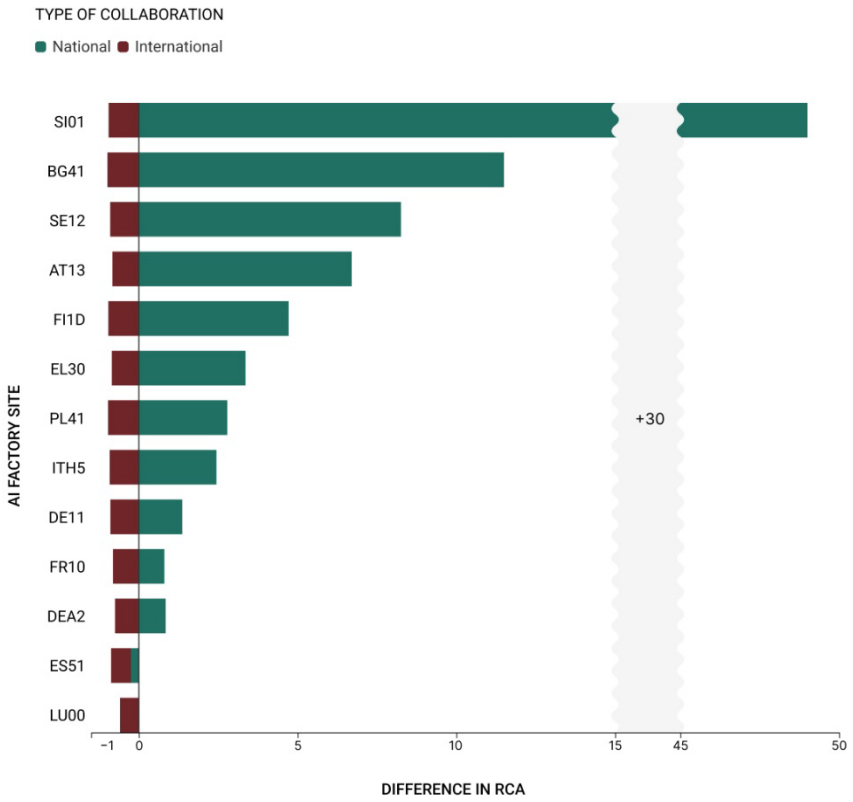


Source: CEPS

Notes: same metric as in Figure 6. See interactive version *here*.

Our analysis suggests that the patterns of collaboration among the AI factory sites are weak, and certainly weaker than those of the leading regions in AI-related patents. While existing collaborative efforts were not included in the selection criteria for the sites, announced plans for collaboration were one of the requirements for selecting sites. Our findings suggest that the implementation of these plans will be vital for the success of the AI factories, especially given their relatively small size, which in practice will require cooperation and networking between sites (and with gigafactories). Otherwise, the sites may remain essentially national hubs, attracting national talent, given that they collaborate mostly within their borders (Figure 8). This is particularly concerning given that the US AI innovation network appears to be considerably more integrated (Balland et al., 2025), placing collaboration between the factories at the heart of the competitiveness problem.

Figure 8. Average difference in the RCA of actual and expected collaborations between AI factory sites and all other European regions



Source: CEPS

Notes: the average difference in RCA of actual and expected collaborations in AI patents between the AI factory sites and the remaining European regions, is shown by collaboration type (national or international). See interactive version *here*.

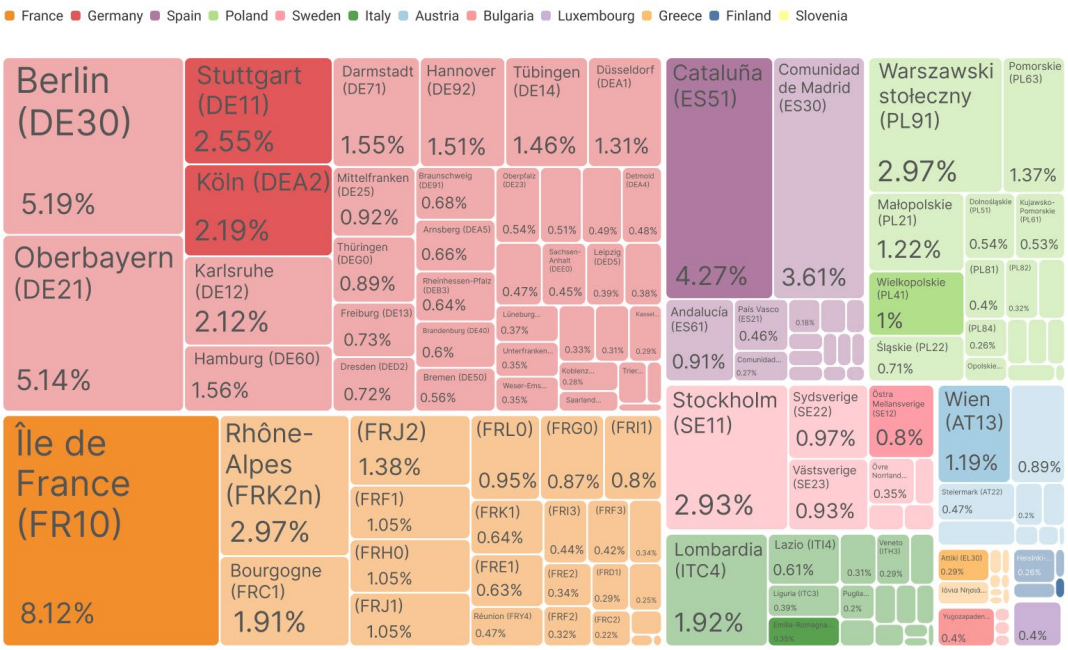
2.4. TALENT ATTRACTION IN AI FACTORIES AND EXCELLENCE HUBS

Another important issue is whether AI factories will be able to attract talent. As explained above, this may not be needed if one assumes that researchers can access compute infrastructure virtually. That said, the attraction of talent to the factories has been a recurring argument in the European Commission's statements on the AI continent agenda. The European Commission has framed AI factories as 'dynamic ecosystems that (…) integrate AI-optimised supercomputers, large data resources, programming and training facilities, and human capital'. The Commission's AI in Science strategy also includes measures to attract global scientific talent and secure researchers' access to AI gigafactories.

Figure 9 shows the number of vacancies requiring the most advanced AI skills (as defined in Nurski et al., 2025) for 2023 and 2024 in each region in countries with AI factories. The darker-shaded regions are those hosting factories. In six cases, the highest demand for advanced AI skills is found in regions where an AI factory will be hosted. This is the case for Île-de-France (FR), Cataluña (ES), Wien (AT), Yugozapaden (Bulgaria), Attiki (Greece) and Luxembourg (LU) (which only has one region).

Among these cases, only Île-de-France and Cataluña have significant demand for advanced AI skills in absolute terms (Figure 10). The hosting regions of Köln and Stuttgart lag behind Berlin and Oberbayern in Germany, though their numbers are still significant (they are among the leading 20 regions in Europe for the total number of vacancies). The fact that many of these regions exhibit the highest demand within national borders underscores the risk that the AI factories will remain national hubs.

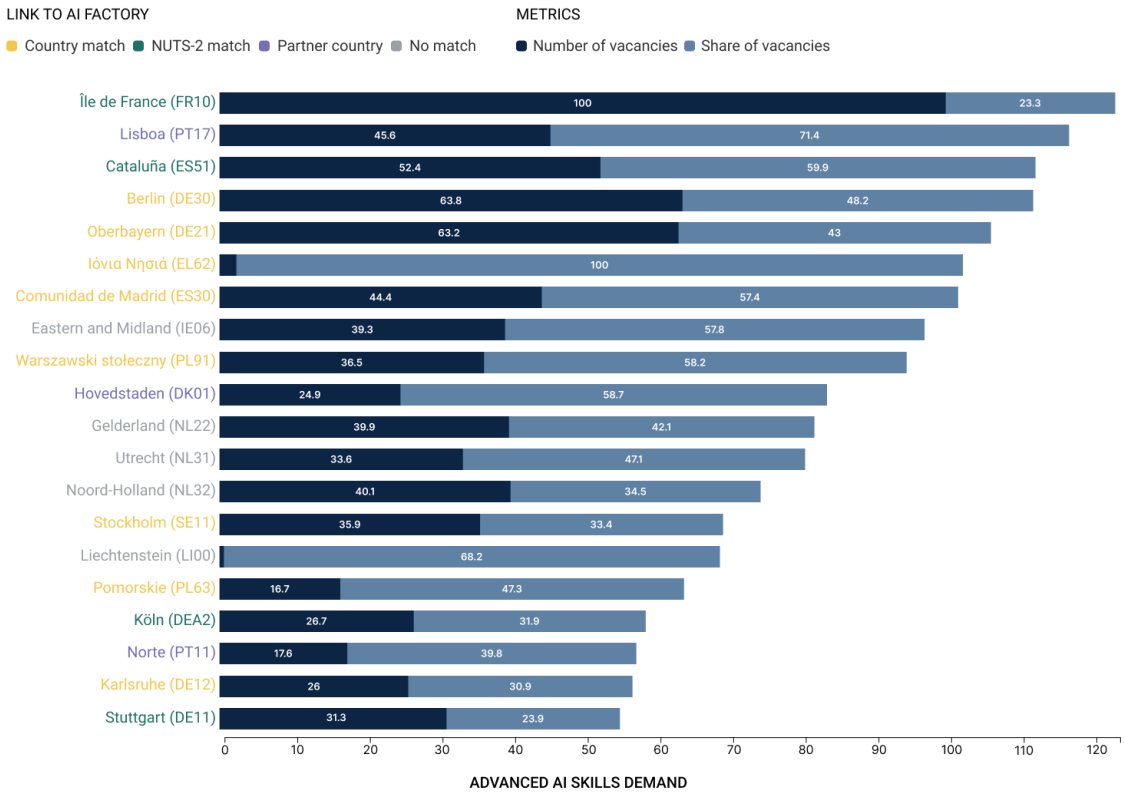
Figure 9. Number of vacancies seeking advanced AI skills per region in countries to host AI factories



Source: CEPS

Notes: Number of vacancies (2023–2024) seeking advanced AI skills per region in each country with an AI factory. Regions with fewer than 10 vacancies for the period are excluded from the analysis. The regions in darker shades, compared with the rest, are those hosting AI factories (two for Germany and one each for the rest of the countries). See interactive version *here*.

Figure 10. Top 20 European regions based on the number of vacancies seeking advanced AI skills and share of all vacancies



Source: CEPS

Notes: the AI vacancies index combines the (normalised) number of vacancies seeking advanced AI skills and their share of all the region's vacancies in 2023–2024. Regions with fewer than 10 vacancies for the period are excluded from the analysis. The names of the regions are colour-coded to denote any link with a factory: green is a match with a factory; orange is a region in a country that hosts a factory; purple is a region in a country that is part of an AI factory consortium; and black is none of these.See interactive version *here*.

Considering Europe as a whole, not attracting or retaining talent where there is most demand for advanced industrial R&D in AI might, on one hand, hinder these regions' ambitions for AI innovation, and on the other hand, create an AI talent surplus in the hosting regions. Of course, this view does not account for the other side of the equation – the current levels of talent across European regions – nor does it consider future increases in demand if investment accrues to the factory sites. In any case, the results suggest a degree of caution about the potential talent attraction of the AI factories and may also provide insights for the selection of sites for future gigafactories.

Another important aspect to consider is the private sector's considerably stronger ability to attract talent, due to higher salaries and working conditions. The projects for private compute infrastructure mentioned in Table 1 above, if not sufficiently combined with

public AI factory investment, may end up catalysing most of the stock of advanced AI experts and researchers.

2.5. CAN AI FACTORIES ACTUALLY COOPERATE?

At present, EuroHPC supercomputers are accessed through EuroHPC access calls governed by specific policies and requirements, but they do not yet possess fully federated capabilities. According to the AI Continent Action Plan, the EuroHPC Joint Undertaking 'will serve as the single entry point for users across the EU, providing access to computing time and support services offered by any EuroHPC AI Factory'.

To advance this goal, the EuroHPC Joint Undertaking signed a contract at the end of last year with a consortium led by CSC–IT to develop a federation platform designed to realise this vision. The platform will provide a unified portal for discovering and requesting compute and data resources across all EuroHPC supercomputers – including AI factories and gigafactories. It will enable harmonised identity management, authentication, and cross-job dispatch for multi-site workflows. Its successful implementation will be essential to ensuring seamless remote access to these facilities and fostering infrastructure integration and collaboration among users across different factories. However, the project is likely to face substantial challenges in implementation.

Built for academic supercomputing use cases, European HPCs rely on very different stacks, made of often incompatible hardware-software suites. As observed by Segler (2025), developers building models in Finland's LUMI, the only one using the AMD ROCm software stack, would have significant problems in porting their code to Nvidia-using HPCs, such as Spain's MareNostrum 5 or Italy's Leonardo. Even within factories using the Nvidia full-stack ecosystem, different accelerators pose interoperability problems: Leonardo, MareNostrum and the smaller factories in Luxembourg and Slovenia use Nvidia H100, whereas JUPITER and Bulgaria's Discoverer rely on Nvidia H200. The same fragmentation may be observed for all complementary hardware and software, creating hurdles for portability and ease of cooperation at scale.

This is, of course, no surprise. Large hyperscalers such as Amazon, Microsoft, Google, and now Tesla, OpenAI and Nvidia, all use a vertically integrated, seamlessly interoperable, privately assembled technology stack to power their factories. In particular, many of the tech giants rely on Nvidia, often coupling it with in-house special-purpose accelerators (e.g. Tesla's DoJo or Meta's MTIA), whereas Google uses its own tensor processing units (TPUs) optimised for AI, increasingly challenging Nvidia's current leadership in GPUs. Like in personal computing, this provides enormous advantages for scientific and technical cooperation. Against this background, the EU's approach to scaling up HPCs into AI

factories may face major obstacles, not only for technical interoperability, but also for the portability of skills and expertise across factories.

All in all, this problem should be duly considered if the EU plans to use AI factories as nodes of a pan-European network. It will not suffice to enable cloud-based remote access to facilities detached from hubs of excellence; the network will also need to be genuinely federated. This holds especially if, once the EU realises the very small scale of these infrastructures, it plans to use them as antennas of a multi-tier network orchestrated by (currently Nvidia-dominated) gigafactories.

3. THE GIGAFACTORY MODEL: WILL IT WORK?

Compared with the AI factory model, the future gigafactories announced by von der Leyen at the Paris AI Action Summit should feature more complex governance, relying on a public–private partnership where 35% of the capital is public, possibly de-risking private investment, and thereby attracting co-funding by private investors (all operational expenditure will be private).

Up to five gigafactories, each equipped with at least 100 000 high-end accelerators, would bring respectable scale to the EU. This is so even if (as already mentioned) the size of the investment would remain much smaller than the figures hitting the news every day in the US, and may be dwarfed by the sums that could emerge from the current wave of infrastructure investment. In other words, even if the EU accelerates on investment, other countries are certainly not standing still, and are likely to move faster. AI gigafactories will replicate the scale of similar projects announced mostly by Nvidia over the past few weeks (see Table 1), with CEO Jensen Huang rushing across the globe to secure fast adoption of its 'sovereign AI' offer. Not surprisingly, Nvidia has taken a leading role in shaping the gigafactory model, presenting full-stack solutions that transcend hardware and extend to the cloud and future robotics solutions.

Apart from the funding model, a central difference between factories and gigafactories is that the former will mostly be reserved for research and industrial AI R&D, while the gigafactories will target large-scale inference. Moreover, paid commercial access is reserved for only 20% of EuroHPC compute (or 10% of the total compute), while it will be more than 65% for the AI gigafactories (though note that it has not yet been disclosed whether all the access managed by the AI gigafactories coordinator will be pay-per-use).

The application process for the EuroHPC's free access mode for the AI factories involves peer review for both research and industrial innovation projects; access modalities and procedures for the AI gigafactories remain to be established. To be effective in supporting industry, these would benefit from providing more flexible usage than the project-based, time-bound application process of the AI factories. Current EuroHPC regulation places an emphasis on trustworthy AI in reference to the AI factories, while it stresses security and energy efficiency with respect to the operation of the gigafactories. But all three elements are central to the intended impact of the initiatives on European competitiveness.

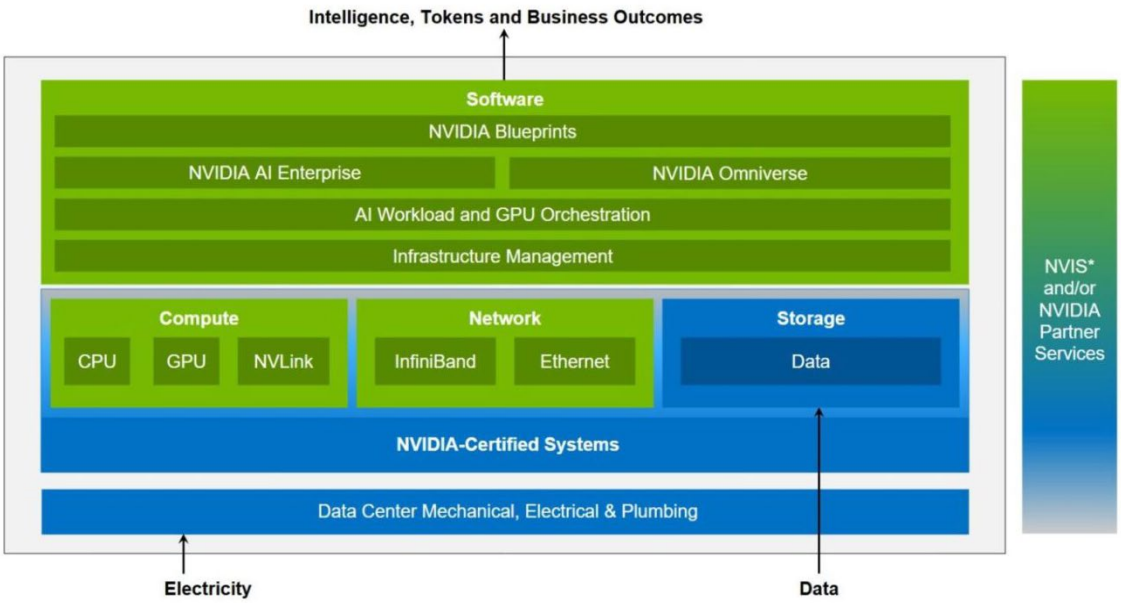
The EU initiative, now nested in the AI Continent Action Plan, met with strong interest among investors, with 76 expressions of interests received in the course of a few weeks. The Commission announced an investment of EUR 20 bn (repurposed from existing programmes such as the Digital Europe Programme, Horizon Europe, and InvestEU), and has allowed Member States to earmark cohesion funds to double the public share.

A central rationale for the AI gigafactory initiative is to provide Europe with the ability to train and host state-of-the-art AI models on European soil. At present, only a handful of private actors (mostly US technology firms) possess the combination of hardware scale, proprietary data, and capital required for this task. European research organisations and start-ups typically depend on foreign cloud providers or negotiate ad hoc access to national supercomputers whose architectures were not optimised for AI workloads. The gigafactories aim to remove this structural barrier by establishing shared facilities in which the most compute-intensive phases of model development can occur under European jurisdiction.

That said, training modern, large language or multimodal models entails enormous computational and organisational complexity. The hardware side is only part of the challenge. It must be paired with high-bandwidth storage, advanced networking, orchestration software, and human expertise to distribute training efficiently across tens of thousands of accelerators. The data pipelines, pre-processing tools, and evaluation frameworks used to monitor these models also require an infrastructure and governance regime different from traditional HPC. Unless these layers are fully integrated, raw GPU numbers will not translate into genuine training capability. For this reason, the European gigafactory model must be conceived not merely as a hardware deployment but as the construction of an entire training ecosystem, encompassing compute, data management, model evaluation, and safety testing.

Figure 11 below exemplifies the technology stack of an AI factory, in the version provided by Nvidia, likely to be by far the dominant supplier of GPUs for AI gigafactories. As shown, Nvidia is much more than a hardware producer: it couples its GPUs with a long list of additional hardware and software solutions, some of which are central to the operation of the AI factory. Among them, one important component is Compute Unified Device Architecture (CUDA), a set of proprietary tools and an Application Programming Interface (API) that runs on top of an operating system like Windows, Linux or MacOS. It provides users with the software (including a compiler, libraries, and developer tools) that enables applications to run on Nvidia GPUs. CUDA gives developers a way to harness the parallel processing power of GPUs for tasks beyond graphics.

Figure 11. Nvidia's full-stack offer for AI factories



Source: Nvidia.

Given that gigafactories are expected to operate as public-interest infrastructures, accessible to academia, industry, and start-ups, the nature of the models developed within them will determine whether the facilities truly advance Europe's goal of digital sovereignty. In contrast to the AI factories, which have only 10% of the compute reserved for private use under the EuroHPC calls, more than 65% of AI gigafactory compute will be dedicated to commercial use. If the outcome is a small number of proprietary models operated by private partners, the public investment may end up enforcing dependence on a limited group of actors.

However, if a portion of the compute is earmarked for open-weight or fully open-source models, Europe could build a corpus of publicly available foundational models – linguistic, scientific, or domain-specific – that become shared building blocks for the continent's digital economy. Several European initiatives, such as BLOOM or Falcon, have demonstrated the feasibility of training large open models when compute and funding are pooled. Embedding this principle in the governance of the gigafactories would magnify their collective value and distinguish the European approach from the closed, corporate model dominating elsewhere.

The discussion on openness is also linked to data governance. As explained in more detail in the Apply AI strategy adopted in October 2025, the European Commission plans to leverage data spaces to support the development of frontier AI models and AI agents adapted to sectors such as health or manufacturing. For example, the Commission announced its plan to 'facilitate data pooling across industrial actors through trusted third

parties, to ensure a sufficient volume of training data, while preserving intellectual property and data security and making use, as relevant, of the data labs in AI factories'. Gigafactories could serve as trusted data environments in which sensitive or proprietary European datasets could be used for AI training under clear legal conditions, combining technical safeguards (secure enclaves and audit trails) with the protections of EU law. Doing so would allow Europe to exploit one of its comparative advantages – its wealth of structured industrial and public sector data – without compromising privacy or ethical standards.

The gigafactories funded through InvestAI will come from the largest public–private partnership in the world for the development of trustworthy AI to date. The proposed financial structure represents an important innovation. Unlike traditional research infrastructure funded entirely by grants, the InvestAI and EuroHPC frameworks envisage a mix of grants and equity or quasi-equity instruments. Public funds are expected to cover the initial capital expenditure required to build the facilities and secure baseline capacity for research and public missions, while private investors would take equity stakes to expand capacity and commercial usage. This hybrid model could accelerate deployment, attract additional capital, and ensure that at least part of the infrastructure operates on commercial principles, reducing the burden on public budgets.

Yet this arrangement also raises questions about control, access, and accountability. If private partners hold significant equity, they may influence the allocation of compute or the selection of projects, potentially prioritising profitable workloads over open scientific research. Equity participation by hardware vendors might bring technical expertise but also reinforce vendor lock-in if the partners are also exclusive suppliers of chips or software. To avoid this, the investment contracts should stipulate interoperability and non-exclusivity clauses, ensuring that no single supplier dictates the architecture of the facility. Likewise, public grant components should prioritise features that enhance long-term adaptability in the form of modular design, open software interfaces, and the capacity to host multiple types of accelerators.

3.1. WHAT KINDS OF GIGAFACTORIES? FACTORS TO CONSIDER ON THE WAY TO EUROPEAN SOVEREIGN AI

The concept of the AI gigafactory not only raises questions about feasibility and timing, but also compels reflection on what kind of infrastructure Europe wants to build and how it should be governed. Designing such facilities will require choices that extend far beyond hardware procurement. They will determine how Europe balances energy efficiency with scale, public and private roles, talent formation, industrial adoption, research orientation, and ultimately, the very meaning of technological sovereignty in AI.

A first and most tangible factor is energy. The global electricity demand for AI computing is projected to increase more than tenfold between 2023 and 2030, a trajectory that could make AI one of the fastest-growing consumers of power in the digital economy. Each gigafactory will consume the equivalent electricity of a medium-sized city, as training large models involves simultaneous operation of tens of thousands of GPUs, vast cooling systems, and data storage arrays.

To reconcile this with the EU Green Deal objectives, gigafactories must be located in regions that combine low-carbon energy abundance with high-efficiency cooling. They must also rely on additional renewable generation rather than diverting existing capacity. In this respect, northern Europe's energy mix and grid stability offer clear advantages, but the policy framework must also ensure that benefits are distributed fairly and that the overall carbon footprint of the AI infrastructure remains within the EU's climate targets. The proposed Cloud and AI Development Act could provide the legal basis and arguments for requiring these installations to operate exclusively on verifiable renewable sources.

A second consideration concerns the public–private partnership (PPP) model through which the gigafactories are expected to be financed and managed. Europe already has a relatively high share of publicly owned or publicly supported supercomputers compared with the US (Epoch AI). The challenge is to determine the optimal degree of public involvement in the next generation of infrastructure. If public ownership dominates, the system may lack the agility and capital injection needed to keep pace with private competitors; if private investors set the direction, the public-interest mandate risks dilution. The appropriate balance is likely to involve publicly guaranteed baseline capacity for research and strategic projects, complemented by private investment that drives scale and efficiency.

However, this mix demands stringent governance of the EuroHPC-managed compute network to avoid conflicts of interest and to ensure that public access commitments are honoured. Reducing the public share should not mean abandoning oversight; rather, it should mean creating incentives for private capital to invest in AI infrastructure that aligns with European standards and values.

The third dimension is talent. The shortage of top-tier AI expertise in Europe is widely documented. Across the EU, demand for highly specialised AI skills already exceeds supply, and competition for qualified researchers is intensifying (Nurski et al., 2025). Gigafactories, however well equipped, cannot fulfil their mission without the human capital to operate, maintain, and exploit them. This implies significant parallel investment in education and training, from advanced PhD programmes in machine learning and distributed computing to vocational curricula for data engineering and system

administration. Each new facility should therefore be coupled with training pipelines and mobility schemes that attract and retain global talent. Without such measures, the EU risks building infrastructures that depend on expertise imported from the very regions whose technological dominance it seeks to counterbalance.

The fourth issue relates to adoption. While the debate around gigafactories tends to focus on development leadership and the capacity to train frontier models, European firms remain slow to integrate even existing AI capabilities. In 2024, only 13.5% of European enterprises reported using AI technologies, well below the Digital Decade target of 75% by 2030. Adoption remains heavily skewed: about 41% among large corporations, but only 11% to 21% among SMEs. Unless this gap narrows, the additional compute capacity generated by gigafactories will not automatically translate into broader economic competitiveness.

The relationship between training and inference also deserves careful consideration. The gigafactories are primarily conceived as sites for training foundation models – an energy-intensive process that occurs episodically and at massive scale. By contrast, inference – the deployment of trained models for end-user tasks – requires distributed, low-latency infrastructure closer to where data are generated. Building all inference capacity within the same large centres could undermine efficiency and resilience. Europe may need to design a complementary layer of smaller, geographically distributed data centres optimised for inference, while ensuring that model updates and retraining cycles remain synchronised with the central facilities.

Balancing these two functions – training for scale and inference for accessibility – will be crucial to achieving both competitiveness and strategic autonomy. This may be a blueprint for a future division of tasks between gigafactories and factories, provided that the latter are optimised (also) for inference. Such an addition was also suggested by OpenAI in its EU Economic Blueprint, but with no reference to open-source solutions.

Closely connected to these technical aspects is the question of trustworthiness. Europe's AI strategy is grounded in its regulatory commitment to safety, transparency, and human-centric design. Yet the current GenAI paradigm presents unresolved challenges for aligning system behaviour with these principles. The EuroHPC Joint Undertaking amending Regulation adopted in 2024 requires that publicly funded compute be used to develop 'trustworthy and ethical' AI models, but this obligation has not yet been translated into concrete research or evaluation programmes. If the gigafactories are to embody Europe's values, they should be coupled with dedicated research tracks on AI safety, reliability, and alignment, integrating these objectives into the technical architecture of the facilities themselves. Without such coupling, the credibility of

Europe's 'trustworthy AI' agenda risks being undermined by the very infrastructures meant to advance it.

3.2. FUTURE-PROOFING AI FACTORIES? MEMORIES OF THE PAST AND THE FUTURE OF MEMORY

One of the key issues when planning a long-term investment such as AI gigafactories is ensuring that the underlying technological solutions do not rapidly become outdated. Among others, Australia famously had to roll back and reconsider its national broadband infrastructure several years ago, after realising that wireless broadband would have achieved the same universal access goals at a tiny fraction of the cost. Similarly, the extreme dependence of AI factories on Nvidia GPUs, a clear strength if one looks at it today, may become a vulnerability when factories eventually become a reality.

In order to understand why, it is important to look at ongoing developments in the AI infrastructure market. For this, it is important to analyse technology architectures and stacks as biologists study the evolution of life. Our brain is the result of hundreds of millions of years of natural selection and gradual sophistication, with areas of the brain specialising in specific functions. For instance, the occipital lobe specialises in vision, the temporal lobe in auditory processing and language, the hippocampus in memory formation, the cerebellum for fine motor control, etc. This provides a formidably orchestrated variety of performance, endurance and energy consumption (the brain runs on approximately 20 watts of energy).

In personal computing, the need to perform a variety of specialised functions has already triggered a similar transition, not over millions of years but over a decade. Computer processing units (CPUs) remain the reference 'workhorse' for ordinary functions, yet more expensive GPUs have become essential not only for graphics rendering, but also for sophisticated, parallel workloads like AI, simulation, and video. This specialisation is being further deepened with function-specific processors (e.g. neural processing units or tensor processing units for AI and AI accelerators), triggering the need for additional hardware/software in charge of orchestration and integration (motherboard buses, networking chips, etc.).

In all this, a key role in compute AI is played by memory. There too, the need to optimise energy consumption and performance has led, since the dawn of computer science, to a split between short-term and long-term memory, enabling CPUs to access information when necessary, and storing it for slower retrieval when not immediately needed.

In AI, the sheer amount of information that is being processed by CPUs, GPUs and other hardware has required so much performance in memory that related, specialised chips have become as essential as processing units. As a matter of fact, GPUs and AI

accelerators are not just compute-bound, they are memory-bound, since training large AI models requires feeding petabytes of data into processors fast enough to keep thousands of cores busy. Thus, processing and memory are currently co-evolving, and mutually dependent: an advanced Nvidia chip (e.g. H100) depends on a specific type of high-bandwidth memory (HBM3), provided by Korean and Japanese players such as Samsung, SK Hynix and Micron. In the near future, the move to HBM4 will trigger a race as fierce as the one for the next processing unit, considering that memory already accounts for 30-40% of the energy consumption of a data centre.

The coming decade will likely see fast developments in processing, as well as an ongoing convergence between processors and memory, with profound geopolitical implications. While the US dominates processing units, South Korea and Japan dominate memory. Europe, unfortunately, is not a player in any of these areas, despite recent efforts to scale up investment in Italy. Micron has a presence in Italy, and Germany has long courted Intel for logic fabs, but the latest news implies that Intel is entering into an equity partnership with Nvidia, with a strong role played by the US administration. As intelligence spreads into more decentralised infrastructure (edge computing) and eventually connected objects, the frontier of compute and memory will increasingly require constant innovation and agility.

In other words, the next decade of chips will not be determined solely by Nvidia's GPUs or Google's TPUs, but by the ability of nations and companies to secure memory technology and integrate it with compute. As compute and memory converge, the geopolitical stakes will converge as well. Control over memory chips, once seen as a commodity, will increasingly determine technological sovereignty. The future of AI, cloud computing, and even national security will depend not only on who makes the fastest processors, but also on who controls the flow of data that fuels them.

Against this background, several issues must be considered in the quest for EU technological competitiveness and sovereignty.

First, betting on one technology stack for AI factories is unlikely to do any good for the EU's sovereignty claims, especially if that stack is linked to the dominant provider of the time (Nvidia) and may lock in European customers and expose them to extreme dependency in the future. In the coming years, as generative AI becomes increasingly personalised and incremental advances in efficiency and accuracy plateau, new paradigms are likely to emerge, possibly bringing the competitive advantage back into the hands of large-scale 'retail' giants such as Microsoft, Google, Meta and Amazon, with their large customer-installed bases and loyalty.

Co-opetition between these players is likely to resume and reach new levels. For example, Nvidia is trying to invest in AI and diversify its solutions to become less dependent on its partners; Google is building TPUs to outcompete Nvidia on its integrated cloud offer; Microsoft, Google and Meta are investing in alternative AI solutions (e.g. Anthropic and Gemini) to reduce their dependence on OpenAI; OpenAI has reached agreements with Oracle in a recent USD 300 bn deal, inter alia to reduce its dependence on Nvidia.

China recently decided not to buy Nvidia chips, evidently counting on domestic supply to have reached a sufficient level of sophistication and performance. The US is likely to continue leading on AI accelerator design (Nvidia, AMD, Google, Apple and Tesla), but without a secure memory supply, its dominance is fragile. This can also explain other geopolitical trends, such as why Washington is pushing for joint US–Japan–Korea semiconductor alliances.

Second, besides the current supply of advanced processing units, the EU should nurture its relationships with partner countries such as South Korea and Japan to ensure mutual benefits in building a technology stack that can handle advanced memory processors and couple them with a variety of processing units and AI accelerators. Japan has long invested in novel memory technologies (MRAM and ReRAM) through companies like Renesas and Kioxia. These could give it a disproportionate advantage if storage-class memory or persistent in-memory compute takes off. And South Korea will probably remain the DRAM powerhouse but is likely to gradually come under pressure from the US and China. On the other hand, Europe is not in a leading position in memory chips but can offer market opportunities and complementary technologies that would need to be sufficiently mapped.

Third, unless specific safeguards are in place, betting on the existing model of gigafactories may mean, for the EU, giving up its ambition to adopt a vision of trustworthy AI, which is at once human-centric, sustainable and resilient. Placing faith in the Nvidia-led architecture is likely to replicate the current energy- and water-hungry approach to generative AI, and betting on an approach to GenAI that is falling short of the extraordinary achievements it repeatedly promised. For example, China has shaken the market with its ability to develop cheaper, open-weight, very powerful AI models such as DeepSeek R1. They rely on a less costly and more energy-efficient, non-cutting-edge set of (approximately 2 000) Nvidia GPUs, partly due to the export controls imposed by the US that made the more advanced Nvidia chips unavailable.

The recent decision by China to restrict purchases and step up customs enforcement on Nvidia AI chips suggests that the country is close to developing a competing fleet of homegrown chips. These include Huawei's CloudMatrix system (which attains a similar performance to Nvidia's by linking together smaller processors) and in-house chips

developed by Alibaba and Tencent. And even on the memory chips side, while trailing behind South Korea and Japan, China is ahead of the US and seems poised to increase its current market share (5%). One company, Changxin Memory, reportedly has the capacity to cover 13% of the global market.

In summary, the EU should look way beyond GPUs (and Nvidia) when defining its AI strategy. It should rather learn from China than the US. And it should partner with South Korea and Japan to lay the foundations of future AI modes and applications that combine world-class memory with powerful computing capacity optimised for AI. And it should tailor AI applications to specific industry needs, rather than simply chasing the dream of artificial general intelligence.

One of the avenues to be pursued, possibly through a 'moonshot' at the EU level, is the development of more energy-efficient, reasoning-oriented chips known as 'neuromorphic' chips. Alternative, more explainable and traceable approaches, known as neuro-symbolic AI, incorporate symbolic reasoning, allowing for the explicit representation of knowledge and rules. These approaches have been subject to extensive research, including in Europe. They promise to adhere more closely to Europe's vision of trustworthy AI, which entails inter alia energy efficiency and the traceability of decision-making processes. On this, practices such as the back-propagation of spiking neural networks could become important for addressing outstanding regulatory challenges.

On neuromorphic computing, Europe can count on a past moonshot that may create significant spillover effects: the EBRAINS platform, the successor to the Human Brain Project, allows researchers to remotely run spiking networks on two complementary, European neuromorphic systems: SpiNNaker (digital and real-time) and BrainScaleS (analogue and accelerated). Germany's SpiNNaker2/SpiNNcloud effort (TU Dresden plus SpiNNcloud Systems) is moving from research chips to multi-board systems and even supercomputer-class deployments, with fresh EU innovation funding and public milestones. The role of world-leading research centres such as imec, in Leuven (Belgium), can provide a further boost to the ecosystem, which increasingly counts important players in vision (Prophesee and iniVation) and hardware (SynSense and CEA-Leti/Heidelberg).

At the same time, China's new Darwin Monkey, launched in August 2025, outcompetes European neuromorphic chips by including as many as 2 billion artificial neurons and over 100 billion synapses. This must be compared with the 158-180 million neurons reached by SpiNNnaker/SpiNNcloud). These developments have not gone unnoticed in the US: OpenAI signed a letter of intent to purchase USD 51 mn of neuromorphic chips from Rain AI in December 2023, to be deployed on inference. The recent USD 5 bn partnership

between Nvidia and Intel, announced in September 2025, may include work on some of the world's most advanced neuromorphic chips (Intel's Loihi 2).

3.3. BUILDING NEW DEPENDENCIES RATHER THAN PROMOTING TECH SOVEREIGNTY?

A very important aspect of the EU's tech sovereignty agenda is related to the potential impact of proposed investments on Europe's current dependency on foreign vendors. It will not have gone unnoticed, from the analysis of Table 1, that the overwhelming majority of GPUs installed in Europe are provided by Nvidia, currently the dominant vendor in the field, and certainly not a European company. This seems a recurrent pattern also outside the EU, for example in the UK and in the UAE, but also in Taiwan, Malaysia and several other countries, where Nvidia partners with national governments and/or large companies (e.g. telecom operators) to deploy its own concept of an AI factory.

This choice prompts at least four overarching questions. First, could the EU remain tied to a suboptimal vendor if Nvidia is outcompeted by other players in the coming years? Second, what if the current reliance on GPUs becomes a burden as other computer processing technologies prove superior? Third, does reliance on Nvidia GPUs introduce rigidities as to the type of AI that gigafactories will be able to support? Fourth, will excessive reliance on Nvidia undermine the EU's aspiration to build a fully sovereign 'EuroStack'? Below, we explore each of these questions.

3.3.1. *What if Nvidia loses the GPU throne?*

One risk scenario for Europe's current investment in AI gigafactories is related to the mounting competition with the dominant player Nvidia in the construction and deployment of GPUs. Some of this competition is being triggered by dominant AI providers (such as OpenAI), which are trying to avoid becoming too dependent on Nvidia chips. As a result, they are consequently promoting rival players like AMD (just as IBM did with Intel, requiring that it shares the x86 microarchitecture with the same company, AMD) so that they can gradually scale up and improve their production of GPUs. The advantage of AMD GPUs such as MI300/Instinct is that they are run by the ROCm AMD software stack, which unlike Nvidia's CUDA, is open source.

It would be ironic to find out that OpenAI is diversifying to avoid becoming too dependent on Nvidia, but the EU is not doing so. More precisely, some of the AI factories rely on alternative processors, including LUMI in Finland and the Hunter supercomputer in Stuttgart, running on AMD. But the gigafactories appear to be entirely reliant on Nvidia. The same can be said for recent deals at the national level, such as the OpenAI/SAP agreement in Germany, which seems destined to rely on Nvidia infrastructure. This is due to the fact that both companies have a strategic partnership with Nvidia. The OpenAI/Nvidia partnership, announced on 22 September 2025, will deploy at least

10 gigawatts of Nvidia systems for OpenAI's next-generation compute infrastructure. In March 2024, SAP announced that it plans to integrate Nvidia's generative AI and GPU-accelerated capabilities into its cloud/enterprise stack.

Other competitors may gradually erode Nvidia's leadership in the market. In the US, the fast evolution and diversification of computing and AI use cases is leading to the emergence of companies selling specialised products (e.g. inference, narrow domain, wafer-scale, or edge), rather than general-purpose GPU replacement. These rivals may gradually conquer niches or architectural advantages (latency, power, memory bandwidth, and interconnect) where GPUs are less optimal. For example, Cerebras' wafer-scale engine is extremely large (many trillions of transistors) and optimised for massive model training & inference. The company claims very high performance per chip, recently receiving USD 1.1 bn in funding (but later filed to withdraw from its initial public offering process).

The rise of Nvidia competitors from the US, such as AMD or Cerebras, may not represent a significant problem from the standpoint of EU gigafactories and the risk of vendor lock-in. The reason is that in recognition of Nvidia's dominance, these competitors are designing their GPUs to co-exist with Nvidia's, and are compatible with the latter's CUDA software. In other words, future gigafactories may use 80% of Nvidia GPUs with 20% coming from AMD or Cerebras. Many hybrid clusters today run both Nvidia and AMD GPUs under unified, open-source orchestration (e.g. Kubernetes).

But what if rivals from other countries, namely China, outcompete American ones? Chinese companies such as Cambricon, Huawei and Alibaba are already almost at par with Nvidia, and are expected to grow even faster now that the Chinese government has banned the purchase of Nvidia chips by Chinese companies. Should these solutions become more viable on the market, the EU would be faced with the difficult prospect of GPUs that could technically co-exist, but which would practically be hampered by export controls, vetoes, and possible lack of interoperability at the software/runtime level. More specifically, US companies (Nvidia and AMD) are barred from supporting integration with blacklisted Chinese chips, hence mixed clusters would violate export-control or IP-sharing laws. Also, the software stacks of Chinese vendors mimic Nvidia's CUDA, but are not interoperable at runtime, which makes the porting of models possible, but prevents individual tasks from running across both infrastructures.

3.3.2. *What if GPUs are not the (only) future of AI?*

Another issue to consider is whether GPUs, originally designed for graphics and not for machine learning, were gradually replaced by other types of processing units. For example, Google's proprietary TPUs are optimised for AI and are offered exclusively

through Google's managed cloud, with no on-premises option and no open hardware or driver interface. Were European institutions to rely heavily on TPUs for large-scale training, they would cede direct control over both compute resources and compliance oversight. In practice this would place strategic AI capabilities within the operational jurisdiction of a single foreign cloud provider, an arrangement difficult to reconcile with the EU's ambitions for digital sovereignty and resilience.

Moreover, were TPUs to eventually dominate the market, the EU would find itself between a rock and a hard place, having to choose whether to continue using a suboptimal solution (the outcompeted Nvidia GPUs), move to a proprietary full-stack running on a non-EU cloud, or try to orchestrate GPUs and TPUs. The latter is possible but not ideal, as the two processing units have different runtimes and cannot be accelerated or trained together.

Another development that may challenge Europe's current approach is RISC-V, an open, royalty-free, instruction set architecture created at the University of California, Berkeley, and governed today by the Swiss-based RISC-V International Foundation. Unlike proprietary architectures such as ARM or x86, RISC-V allows any company or nation to design its own processors without paying licensing fees or depending on foreign intellectual property.

For China, this openness has become the foundation of a sovereign computing strategy: firms such as Alibaba, StarFive, and the Chinese Academy of Sciences are building RISC-V chips that power everything from embedded devices to early data centre processors. Notably, the RISC-V architecture does not rely on GPUs, but rather on CPUs (the traditional computer 'workhorse') that coordinate workloads, control accelerators, and power billions of embedded and edge devices.

RISC-V cores can scale from tiny microcontrollers in sensors and autonomous vehicles to data centre processors capable of running Linux or cloud workloads. Designers can easily strip away unnecessary functions or add domain-specific extensions for cryptography, vector operations, or AI inference. This flexibility translates directly into energy efficiency: RISC-V chips can be optimised for ultra-low-power tasks, consuming milliwatts instead of the tens or hundreds of watts typical of general-purpose CPUs. For Europe's industrial internet of things (IoT) and cutting-edge AI ambitions – where most energy use comes from vast numbers of distributed devices – such efficiency could yield enormous carbon and cost savings.

For Europe, RISC-V thus represents both an opportunity and a warning. On one hand, it provides a global commons for processor innovation that could help European research institutions and chipmakers regain design sovereignty without relying on US or Chinese

licensing regimes. On the other, Europe's current AI infrastructure depends overwhelmingly on Nvidia's proprietary GPUs and CUDA software ecosystem. Reliance on CUDA may limit Europe's ability to optimise hardware locally or integrate heterogeneous accelerators from domestic vendors.

3.3.3. *Chasing Icarus: will gigafactories constrain Europe's approach to trustworthy AI?*

The natural consequence of the arguments presented in the two previous sections is that Europe may have to think very carefully about its approach to AI, especially if it wants to champion a more energy-efficient, scalable, and modular paradigm. Such an endeavour would enable the deployment of tailored AI solutions across a spectrum of computational needs, from training massive frontier models that demand huge amounts of energy, water, and data, to running smaller, specialised models on modest infrastructure or at the network edge. In this vision, AI would become a distributed and flexible capability, allowing European industries, public services, and researchers to innovate using compute resources that match their scale and sustainability goals. Perhaps this would also address the mounting concerns revolving around the lack of an obvious business case for the AI gigafactories.

Achieving this balance requires rethinking the dominant 'bigger-is-better' mindset currently shaping global AI development, particularly in the US. The EU has a strategic opportunity to define a model of AI growth that values efficiency, interoperability, and trust over brute-force scale. This means nurturing hardware-software co-design around open architectures and low-power computation, not only for climate reasons, but also to reduce strategic dependency on a single vendor or technology stack. Approaches such as DeepSeek's R1 model (demonstrating that competitive performance can be achieved with less-advanced GPUs), RISC-V-based accelerators, and neuromorphic or neuro-symbolic AI systems illustrate viable paths toward this goal. Each of these technologies embodies principles Europe values: openness, modularity, explainability, and resource efficiency.

The key question is whether the EU can integrate these emerging paradigms into its evolving AI infrastructure strategy. Doing so will require aligning industrial policy, research funding, and procurement frameworks. Horizon Europe, the Chips Act, and InvestAI could jointly support testbeds and pilot deployments of these alternative computing models within the planned AI factories. Embedding such diversity early would ensure that Europe's gigafactories become platforms for experimentation and evolution, not fixed monuments to a single generation of hardware. In essence, the EU must design its AI architecture as a living system, capable of adopting new compute paradigms as they

mature, if it is to reconcile competitiveness with sustainability and sovereignty in the long term.

3.3.4. *Could the current strategy undermine the EuroStack?*

Last but not least, it is important to tackle the most straightforward question. Will investing in collaboration with a non-EU company, or a few non-EU companies, undermine the EU's stated ambition to achieve technological sovereignty, and build a fully-fledged EuroStack? Many concerns have recently been raised, especially since SAP announced its agreement with OpenAI in September 2025. Proponents of the EuroStack have reacted heatedly, denouncing the agreement as sovereignty washing. SAP has responded by explaining why the agreement would not violate technological sovereignty, since sovereignty (in SAP's interpretation) is not about reinventing every layer, but about owning the rules, the operations, and the data.

In reality, much depends on how sovereignty is concretely defined. Experts typically distinguish between the following types:

(i) **data sovereignty**, which includes all aspects of information management that are subject to the rules of its originating jurisdiction, regardless of where data is actually located;

(ii) **operational sovereignty**, i.e. the degree to which a customer organisation has visibility into and control over the provider's operations; and

(iii) **technological sovereignty**, i.e. the degree to which a customer organisation can ensure the continuity of and control over its right to technological autonomy (this includes the ability to operate disconnected from a technology provider).

For SAP/OpenAI, data and operational sovereignty are in principle guaranteed, since the infrastructure would be operated by Delos Cloud, located in Germany and operated by 'German citizens' under the control of the German government. That said, it remains to be seen if the 'but for' rule of technological sovereignty (i.e. can the system continue operating if disconnected from the underlying technology provider) is actually guaranteed.

Still, an important question is whether 'sovereignty' should be taken as coinciding or not with 'reinventing', or Europeanising, all layers of the stack. In principle, one could argue that the infrastructure on which AI is run is less important than ownership of, and control over, the higher layers of the stack, such as software orchestration, data, and AI models. Using a wildly oversimplified metaphor, one could argue that a runner who wins an Olympic marathon should not be considered less 'national' if running in foreign shoes or a foreign outfit. Transposed to the AI factories world, this would be tantamount to asking

whether sovereignty can be achieved on a foreign, yet allegedly 'sovereign' infrastructure such as that currently proposed by most American big tech.

A key question to address in order to solve the puzzle is whether the AI (giga)factories being built in Europe will feature reliance on Nvidia CUDA or not. Another is whether Nvidia will be asked to open its architecture to enable orchestration of a variety of solutions on its underlying 'engines' (the GPUs). There are two scenarios that raise concerns: (i) CUDA becomes the de facto industry standard, leading Nvidia to keep control of higher layers of the stack; and (ii) Nvidia agrees to open its CUDA to other solutions, but then in effect limits interoperability so that only Nvidia GPUs with CUDA allow for optimal performance, training and acceleration, including for specific industrial use cases.

These dynamics are not new to the tech world. Back in the early 1990s, the importance of the operating system layer for controlling the whole stack became vividly clear with the rise of Microsoft in the broader IBM-led personal computer stack. Later, the 'Wintel' model became the centre of gravity of the whole evolution of the PC. The Microsoft vs Netscape/Java saga at the end of the 1990s showed the strategic importance of fighting to conquer the critical platform layers of the stack.

The list has only grown longer over the past two decades, with ongoing fights for dominating the most strategic layers within a common stack, rather than the forking of technology stacks. Today, cross-ownership and strategic partnerships between large tech companies (e.g. OpenAI with Intel, Nvidia and AMD; Microsoft with OpenAI; Google and Amazon with Anthropic) portray even more complex dynamics. These may inevitably reverberate on the ability of the EU to rely on a rational business strategy when allowing non-EU players to deploy massive 'infrastructure plus' on its territory. The intricate business relationships between American companies, largely undisturbed by antitrust law on this front, may have to be taken into serious consideration when planning for the future of EU infrastructure.

This does not necessarily mean that 'CUDA is the new MS-DOS'. It simply raises the question, urging EU leaders to impose conditions and plan very carefully when designing gigafactory architectures. This is even more crucial given that they will have to do this in a position of very weak bargaining power. If one interprets bargaining strength, following the Harvard Business School jargon, as BATNA or the 'best alternative to the negotiated agreement' (Ayres and Nalebuff, 1997), it is clear that the EU, as of now, has virtually no alternative to building the factories on non-EU GPUs. Dealing with China might trigger unprecedented retaliation from the US; and notwithstanding the efforts made with the Chips Act (and its recent follow-up), planning for a 'made in Europe' architecture may foster sovereignty, but would unavoidably kill competitiveness.

CONCLUSION: WHAT 'EUROPEAN WAY' TO AI?

This paper has tried to shed light on the complexity of Europe's ambition to regain competitiveness and sovereignty in the AI domain, with specific emphasis on compute infrastructure and its adjacent layers (the technology stack is much thicker, of course). Several important findings arise from our analysis.

First, AI hubs and AI factories are very distinct. The European Commission seems to be following, at least in principle, the same approach as the US in separating the choice of sites for AI factories from the physical locations of the most vibrant AI ecosystems (the regions of London, Paris, Munich, Eindhoven, etc.). Rather than a 'CERN for AI', what emerges is a more distributed network in which factories would benefit from optimal conditions, such as the availability of land, cheap energy and pre-existing high-performance computing. Meanwhile, research ecosystems would continue to grow in other locations, which we call 'hubs' of AI excellence, where research, talent, patents and venture capital create a favourable environment for AI-enabled innovation.

Second, if the assumption above is true, then not all AI factories are being placed in the most ideal locations in terms of energy and infrastructure availability. The ideal conditions for AI factories are mostly in Nordic countries (Sweden and Finland) and far less so in the remaining areas. This must be considered when assessing candidates for the upcoming gigafactories. The EU should avoid building in another source of competitive disadvantage by surrendering to the temptation of seeking 'geographical balance', which is very typical of a complex project like the EU for reasons that are partly understandable. For complex technologies, the economics are clear: innovation is concentrated, and the infrastructure backing requires scale.

Third, although the European investment is significant, it is dwarfed by the levels of investment seen in the US and China. This is even more worrying if one considers speed. The European investment will take time to deploy, and there is no real 'catching up' effect to be expected. While the EU is building the factories with public and private resources, new technological challenges and solutions will appear, and large-scale infrastructure will mushroom in the US and other parts of the world. This leaves the EU with a razor-sharp dilemma: whether to compete where it cannot win, following the lead of trailblazing nations and trying to emulate their model at slower speed and with years of delay; or to chart its own path by developing more human-centric, resilient and sustainable approaches. The latter option would require, among other things, drawing on low-energy alternatives to the current paradigm.

Fourth, while it must clearly procure the most cutting-edge technologies of today, the EU should be careful to avoid adding new dependencies. This requires several safeguards, such as:

- *diversifying sources of GPU supply*, even across the spectrum of possible non-EU vendors;
- *requiring that Nvidia opens up its software layer CUDA and its AI Enterprise suite* to other vendors, under genuinely neutral conditions, to ensure that competition and choice remain open over time;
- *mandating open-source solutions* such as OpenStack on publicly funded infrastructure (e.g. requiring open APIs and container-level portability in AI factory design contracts) to maintain vendor-neutrality and enable multiple solutions for AI acceleration;
- *deepening relations with both South Korea and Japan* regarding the availability of high bandwidth memory chips and collaborative research to prepare for the future era of in-processing memory.

Fifth, the EU should launch a moonshot on AI to explore more trustworthy solutions, starting with the infrastructure layer and its adjacent software layers. Rather than adhering to the dominant, possibly over-hyped wave of generative AI, the EU should ensure that it continues and revives its commitment to more trustworthy approaches. These include neuromorphic AI and neuro-symbolic AI, which may prove to be winning alternatives in the age of physical AI.

The same applies to CPU-based solutions such as RISC-V, which may correspond more directly to Europe's ambitions to deploy open, flexible, modular and scalable approaches to the cloud/edge/IoT. If Europe invests seriously in open CPU architectures, pairing RISC-V cores with European fabrication projects like those under the Chips Act, it could build a new generation of processors optimised for both efficiency and sovereignty. Ignoring that opportunity would leave Europe dependent not only on Nvidia for AI acceleration but also on foreign CPU instruction sets at the heart of every computing device, ceding control over performance, power, and security to others.

All these actions have to be carefully mainstreamed into the modus operandi of the EU when dealing with the AI technology stack. Just like all major tech companies trying to diversify in order to reduce dependencies on their peers, the EU should play the game with the right tactics and strategy, or what we called 'BABE'. For now, that means buying American and (mostly) proprietary solutions. Tomorrow, it means building a future-proof technology stack based on open source, modular, scalable and sovereign solutions. Failing to do so would cripple efforts to achieve a more sovereign EuroStack from the very start, and with it, jeopardise Europe's ambition to chart, together with like-minded countries, a 'third way' towards the AI age.

CEPS
PLACE DU CONGRES 1
B-1000 BRUSSELS