International Center for Future Generations ICFG

CERN for AI

The EU's seat at the table

Authors: Daan Juijn, Bálint Pataki, Alex Petropoulos and Max Reddel

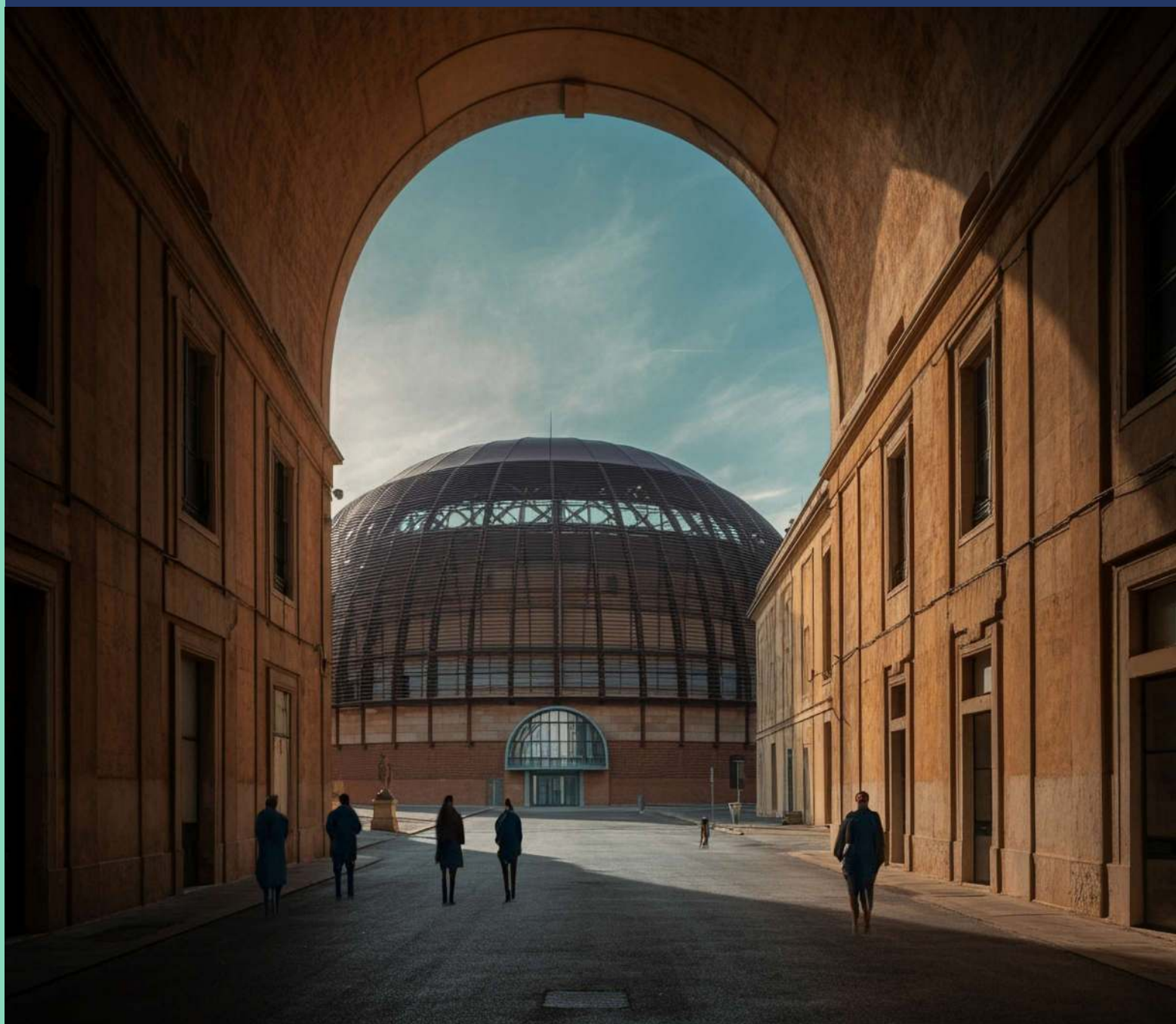

Table of contents

Executive summary

The European Commission President put a CERN for AI at the heart of her vision for addressing the 'hamstrung' competitiveness of Europe. This proposal assesses that the investment needed to compete in the market of increasingly valuable, large-scale, and general-purpose AI models is too large for any single European government or company.

A CERN for AI could boost Europe's economic performance, improve security against external threats, and develop truly trustworthy AI. Europe is lagging behind the US and China in advanced AI and, more generally, tech innovation, mainly because of lower capital deployment and a fragmented ecosystem. A CERN for AI could give Europe the computational infrastructure to build its own frontier AI models, and to spur a thriving ecosystem of high-tech startups and scale-ups, underpinned by talent that would be incentivised to work in and for Europe. Such an ecosystem would benefit not only the private, but also the public sector. A large-scale pan-European effort would further promote the Union's strategic autonomy and enable the development of more trusted, AI-assisted responses to external threats in domains such as cyberwarfare.

Finally, and perhaps most significantly, making frontier AI safe and reliable remains an unsolved scientific problem. The EU cannot gamble on foreign, profit-driven companies to solve this problem, nor can it bank on regulation alone. History has shown that ambitious, European research efforts—like the original CERN—can rapidly expand the scientific frontier. Trustworthy AI can be invented in Europe.

> **"We must now focus our efforts on becoming a global leader in AI innovation. I will propose to set up a European AI Research Council where we can pool all of our resources, similar to the approach taken with CERN."**
>
> **Ursula Von Der Leyen**
> European Commission President

The idea, then, is compelling, but designing and creating an institution like this will require deep planning, strategic allocation of resources, and serious ambition in Brussels and beyond. To succeed, a CERN for AI should have:

- **Multiple paths to trustworthy AI:** Make solving the scientific problem of trustworthy AI its core mission, and tackle it through multiple, targeted, research bets;
- **Competitive compute:** Allocate a budget of €30-35 billion over three years to ensure access to competitive computational infrastructure;
- **World-class leadership:** Appoint leadership that can quickly attract top talent and hit the ground running;
- **Agile and democratic governance:** Balance decisiveness and oversight in the governance structure;
- **Multi-level security:** Strive for openness and transparency where responsible, and security where necessary;
- **Private sector involvement:** Enable the private sector to build upon public, foundational research, and accept their co-funding after rigorous screening;
- **Talent and compute hubs:** Create a single, dedicated talent hub, accompanied by 1-5 separate compute hubs;
- **International partnerships:** Remain open to partnerships with like-minded non-EU countries; and
- **Benefit sharing:** Ensure a benefit-sharing structure among participating governments and businesses.

"The new College of Commissioners should swiftly create an action plan to bring a CERN for AI from dream to reality."

The EU has a unique opportunity to succeed, However, it needs to act quickly: the steps taken in the following months can make or break a CERN for AI. The EU can deliver a Union-wide initiative to pool resources, talent, and ambition into a single, focused effort to develop world-class, trustworthy AI models. To do so, the new College of Commissioners should swiftly create an action plan to bring a CERN for AI from dream to reality.

The first chapter of this report introduces how the idea of pan-European resource pooling for AI in a CERN-like structure gained traction. It also introduces the main pieces of technical and socio-economic context underpinning the paper's recommendations. The subsequent chapter lays out the nine critical features a CERN for AI should have to deliver on its ambitions. The third chapter explains how a CERN for AI could substantially contribute to the EU economy, provide the foundation for resilience-enhancing technologies and steer the trajectory of advanced AI in a trustworthy direction. The fourth chapter concludes the piece.

Introduction

How the idea of a CERN for AI became popular

While the EU leads in the regulation of general-purpose AI models, it has failed to produce a booming domestic industry. Central to this diagnosis is a lack of scale: neither private nor public efforts have equipped European AI researchers with sufficient computational resources and funding. With advanced AI on track to become the defining general-purpose technology of our time, experts are sounding the alarm. More and more people, including Commission President von der Leyen, are calling for a large, publicly funded and centralised institution that puts the EU back on the map. However, such a 'CERN for AI' could take many forms. The details urgently need to be fleshed out.

The last few years saw staggering progress in AI capabilities

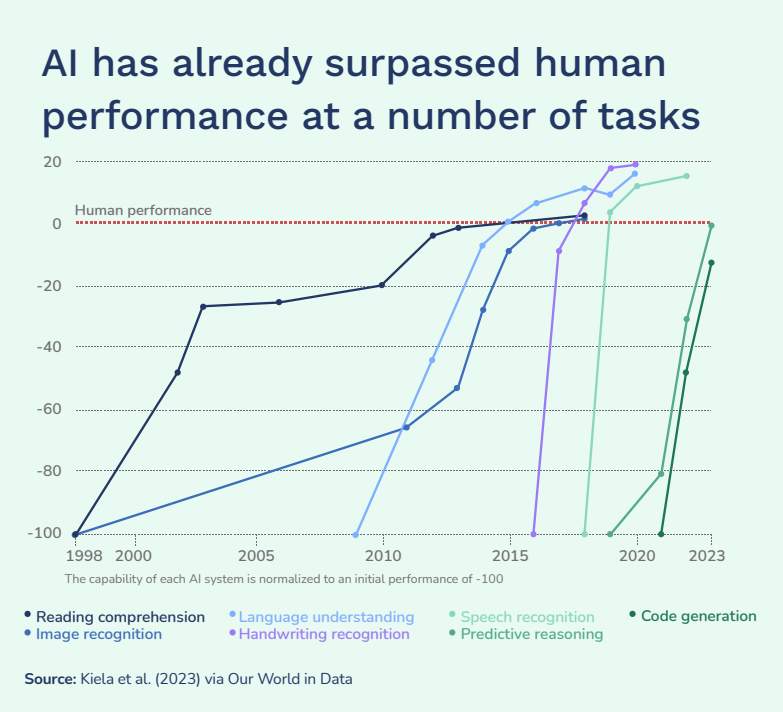

Figure 1: Advanced AI systems are acquiring new capabilities increasingly quickly

Advanced AI is well on its way to become the most important general-purpose technology of our time. In just the past few years, general-purpose AI models have transformed from quirky research projects to productivity-enhancing tools that help millions of users brainstorm, draft reports, or write programming code. Simultaneously, these models have developed a host of new capabilities altogether, like real-time vision, speech indistinguishable from human vocals, and the ability to create photorealistic images. In the near future these capabilities will be condensed into AI assistants—unified agents that not only answer questions in a chat environment, but can help users perform all sorts of real-world tasks in the digital domain. Think of keeping your mailbox up to date, doing your online shopping or creating and

managing a full-fledged website. The most popular of such systems are currently built by OpenAI (ChatGPT), Anthropic (Claude), Google DeepMind (Gemini), Meta (Llama) and xAI (Grok) - all of which are American companies.

Advanced AI refers to highly capable general-purpose models (such as GPT-4, Claude 3.5 Sonnet and Gemini 1.5 Pro) or systems built on top of such general-purpose models. Advanced AI models' capabilities may be confined to language processing, but may additionally encompass 'seeing' and 'hearing', or extend to the generation of pictures, video or audio. In the near future, advanced AI models will likely be given further agentic capabilities.

AI is an exponential technology

The most important driver of this astounding progress in advanced AI has been the exponential increase of computing power – or compute – used to train AI systems. The largest training run in 2023 used approximately 10 billion times more computational operations than the largest in 2010 - similar to the difference between a single human's effort and that of all humanity combined. This relentless increase in computational power enables developers to train bigger models on more data – a recipe that has shown to reliably increase models' capabilities.

Moreover, access to vast amounts of computational power lets developers run more algorithmic experiments in parallel, or create large volumes of high-quality synthetic data, both of which can improve the efficiency of their models. All of this means that AI progress isn't just continuing, it may even be accelerating.

Progress in advanced AI is driven by both increases in computing power and improvements to algorithms

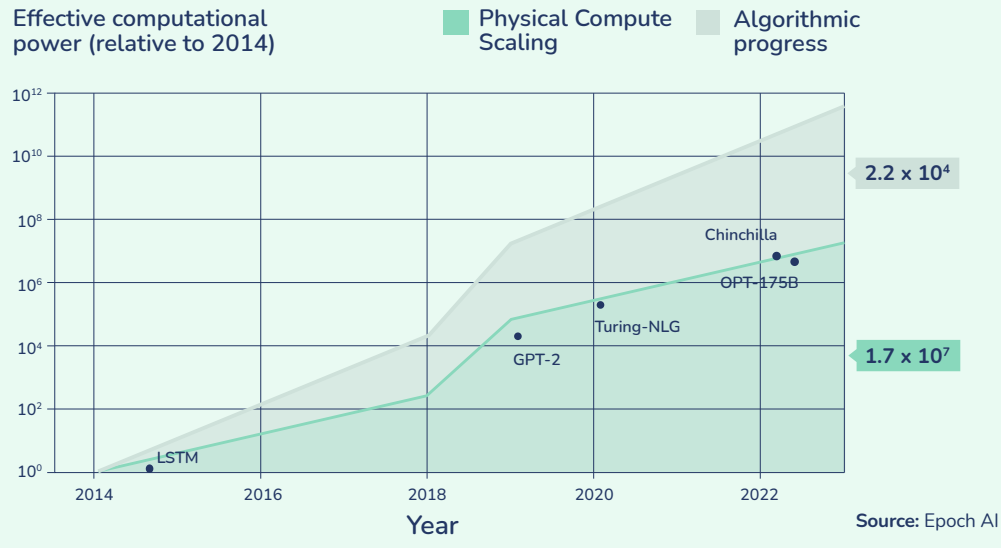

Figure 2 (right): Exponential increases in training compute and algorithmic efficiency have driven the rapid recent progress in advanced AI.

The end of capability progress is not in sight

Computing power and algorithmic innovation will likely remain key drivers of capability progress in the remainder of this decade. On the hardware side, xAI has just started using its new 100,000 GPU cluster and Oracle is reportedly building a 200,000 GPU-cluster housed with NVIDIA's next-generation AI chips, to be used by OpenAI[1]. For comparison, the cluster OpenAI's GPT-4 was trained on only consisted of the equivalent of some 2,000-5,000 of these chips. Meanwhile, Amazon just bought a datacenter location with a dedicated 1 GW nuclear power plant that could provide electricity for a cluster of almost 1 million AI chips. OpenAI and Microsoft are even reported to have begun planning a build-out of a 5 GW AI supercomputer called 'Stargate' that would host 'millions of AI chips'. This massive cluster is supposedly planned to be operational between 2028 and 2030 and could likely be used to train models 10,000x to 100,000x more compute-intensive than GPT-4[2]. At the moment, Europe is not on track to build compute clusters anywhere near the size of those planned in the US.

Next-generation advanced AI could prove transformative

Meanwhile, algorithmic progress seems to be accelerating, rather than slowing down. OpenAI just released GPT-4o mini, a model 100x-200x cheaper than the original GPT-4 released in March 2023, and which seems to be comparable in performance (for OpenAI to be able to produce such an efficient frontier model, they first had to scale up). Although pricing isn't a perfect indicator of algorithmic efficiency, differences this big suggest rapid progress. AI models' reasoning abilities — a skill that many consider crucial to getting AI Agents to work more reliably — are also seeing advances. For instance, Google DeepMind recently debuted AI systems able to score a silver medal in the International Mathematics Olympiad (IMO), only missing gold by 1 point. This target (gold IMO medal) has been a longstanding AI milestone that seemed multiple years away. Now, prediction markets expect it to happen by next year.

The future is inherently hard-to-predict, and progress in advanced AI could theoretically slow down. The market is not betting on it though, with hyperscalers data center expenditures almost doubling over the past year. If progress in AI continues at its current pace, highly capable autonomous AI agents may be developed within the next few years. Such agents could upend European economies by automating labour-intensive tasks and by speeding up science and R&D. If governed irresponsibly, the same systems could also disrupt the job market, lead to large-scale accidents – for instance in financial markets – or be misused by malicious actors to create chemical or biological threats. The stakes are hard to overstate.

[1] Each NVIDIA GB200 'superchip' contains two B200 AI chips.
[2] The models will be on the order of 10^29-10^30 floating-point operations per second (FLOPS) magnitude, contrasting GPT-4's 10^25 FLOPS.

The EU leads in regulation but trails in innovation

The EU's advanced AI industry is far behind

The EU is a global frontrunner in responsible AI regulation. The Union recently concluded the world's first internationally binding AI treaty, the EU AI Act. Such an international agreement on a complex emerging technology like AI is no small feat. At the same time, the EU is dropping the ball on innovation. Europe's Advanced AI industry - i.e. its developers building frontier general-purpose AI models - is struggling to keep up with international competitors, and is at risk of becoming entirely irrelevant. The leading Advanced AI companies are all based in the US, the UK and China, with the very top (OpenAI, Anthropic, Google) concentrated in San Francisco. The best European-made general-purpose model currently scores 11th on the LMSYS arena (a leaderboard that lets users rate the relative performance of general-purpose AI models). The company behind this model—the Paris-based Mistral—is predominantly funded by American VC's and this year signed a controversial deal with Microsoft. Meanwhile, other AI startups inhabiting the same sub-top are struggling: Stability AI leadership is reportedly in talks to sell the company, and both Inflection AI, Adept AI and Character AI were recently swallowed by American big tech entreprises (Microsoft hired 85% of the Inflection AI staff, Amazon 'acquired' Adept AI, and Google did the same with Character AI).

Not only does the EU lack a roster of competitive AI companies, it also lacks the AI computational infrastructure to quickly catch up. All the major cloud service providers that rent out AI chips to AI companies are housed in the US, meaning the EU has become fully reliant on the likes of Amazon, Google and Microsoft for their cloud infrastructure. Through the EuroHPC Joint Undertaking, EU startups do have access to a network of supercomputers, but these machines are not yet equipped for the increasingly large, parallel AI workloads that training frontier AI models requires. Combined, the current eight EuroHPC supercomputers house some 32,000 specialised AI chips, most of which are lower-quality, previous generation NVIDIA chips[3]. Microsoft is reportedly targeting 1,8 million AI chips by the end of 2024, with a much larger percentage of those being state-of-the-art chips. That's roughly a 100x difference in computational resources. Although the EU is investing in expanding the EuroHPC AI clusters through the new AI Factories program, these investments are nowhere near sufficient. Given the distributed nature of the EuroHPC investments, the maximum number of next-generation NVIDIA superchips that

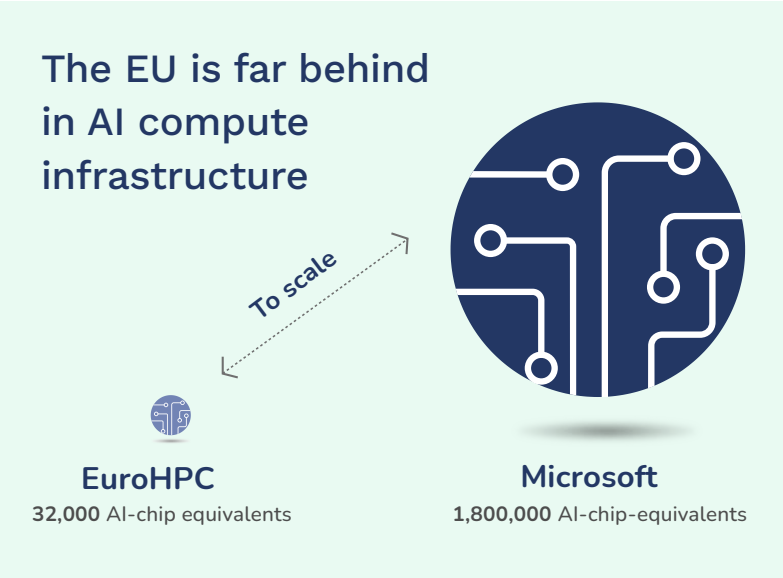

Figure 3: The EU is far behind in AI computational infrastructure.

[3] Lumi houses 11,912 MI250x chips, LEONARDO 13,824 A100 chips, MareNostrum5 4,480 H100 chips, Meluxina 800 A100 chips, Karolina 572 A100 chips and Deucalius 132 A100 chips.

can be acquired by a single EuroHPC supercomputer is capped at some 5,000. For reference: Google has ordered 400,000 of these superchips, alongside millions of its in-house designed TPUs.

In the absence of leading AI companies and necessary computational resources, European AI talent is understandably leaving the continent. Europe delivers 12% of top-tier AI undergraduates, compared to 18% by the US. Yet, partly due to poor retention, this rich pool of talent has not translated into a booming domestic AI industry. Only 71% of AI researchers who went to graduate school in Europe, continue to work in Europe. More than 13% leave for the US. Retention rates are likely even worse at the very extreme, with frontier companies like Google DeepMind offering top-notch researchers millions of dollars in equity to jump ship. Meanwhile AI talent is quickly flowing from the public to the private sector. Governments have trouble securing and retaining technical talent that can help inform policies, while academia cannot offer researchers sufficient compute to stay at universities.

Scale may not be all you need, but sure seems necessary

Why is the EU behind in advanced AI?

There's a popular phrase in modern AI: 'scale is all you need'. Narrowly interpreted, this saying conveys that the quickest way to get something to work in advanced AI is often to just 'throw more computational resources at it'. But a more liberal interpretation also hints at one of the underlying reasons the EU is behind in general-purpose AI. Europe needs more scale, in both public and private efforts. Europe's domestic AI industry suffers from a lack of scale in markets, talent hubs and investment.

Different languages, diverse national regulations and heterogeneous customer preferences imply that it is harder to bring a product to hundreds of millions of users in the EU than it is in the US and China. The fragmentation of the digital single market likely contributes to the often-touted claim that European companies tend to be more risk-averse than their US counterparts: if there is limited potential upside to your business idea, the chance of failure cannot be large. This risk-averse attitude is particularly problematic in venture capital for AI (a segment that is also much smaller in the EU than in the US). Training general-purpose models relies heavily on large, front-loaded investments in computational resources. If funders aren't comfortable with taking large risks, it is hard to attract sufficient seed funding for your AI startup, and thus near-impossible to get off the ground.

Scale is also the missing ingredient when it comes to European AI talent hubs. Although the EU has some solid AI hotspots with leading academics and a lively shell of surrounding businesses, Europe lacks its own Silicon Valley. The agglomeration benefits that come from having a large country's worth of talent and capital concentrated in a single city are hard to overstate. Strong professional networks, limited mobility of talent and world-class support infrastructure have caused San Francisco to attract a whopping 17% of global AI VC funding - more than twice as much as the entire EU.

The EU's public efforts to overcome the limitations of its private sector have so far also fallen short. Although the EU has invested serious public money in general-purpose AI, these investments have been far too scattered. Difficult political compromises have resulted in 'distributed' investments becoming the norm. This diffusion may benefit other sectors, but training a frontier model requires the centralization of talent and computational resources. Programs such as the European ALT-EDIC, in which 16 Member States aim to train language models in their native language, are under-resourced to bring about competitive general-purpose systems. In today's market it is unrealistic to develop competitive AI models with a total budget of under 100 million euros, divided over 16 participants. Access to local data can help address local needs, but is no panacea here: time and time again bigger, multilingual general-purpose models have shown to outperform smaller efforts specializing in a specific language.

Besides the diffuse nature of Europe's public investments, the other salient problem is a limited total budget. The European Commission recently announced that it will invest 800 million euros in new AI infrastructure for the EuroHPC Joint Undertaking over a three-year period. Although this may sound like a lot of money, it is insufficient when contrasted with the 19 billion USD in capital expenditures that Microsoft is spending on data centers every quarter.

CERN for AI: the institution that puts Europe back on the map

Experts are calling for a more unified, large-scale public effort

In light of the EU's poor position in advanced AI, the importance of this new general-purpose technology, and the obvious need for scale and centralization, more and more experts are calling for a large, concentrated European effort to develop trustworthy AI. Common among these proposals are significant price tags (25 billion to 110 billion euros) to achieve the required computing scale, and the recommendation for the EU to invest in fundamental research towards safe and trustworthy AI. Proposals often draw analogies with CERN - the European Organization for Nuclear Research. CERN employs some 70% of all leading particle physicists globally and has become the European billboard for ambitious public efforts after successes like the World Wide Web.

A CERN for AI has recently made its way into the political mainstream, when European Commission President Ursula von der Leyen called for a European AI Research Council in her Political Guidelines. The President describes this as a place where the EU can 'pool all of its resources, similar to the approach taken with CERN'. Considering von der Leyen's proven ability to follow through on major policy initiatives, as seen throughout her first term, the realisation of this AI-focused institution stands on solid ground.

However valid the idea, the proposals for a CERN for AI have left crucial questions unanswered that can make or break this new institution. If the EU is going to invest tens of billions of euros into a moonshot project, it should have more clarity on the objectives, whether the motivations and expected results justify the investments, and how the institute should be designed.

The necessary components of a CERN for AI

SECTION SUMMARY

Moonshot projects only succeed if done right. To deliver on its promises, a CERN for AI needs to pursue multiple research bets with trustworthiness as a north star. To attract top-notch talent in AI, it needs frontier computational infrastructure, and world-class leadership. In order to be competitive, a CERN for AI must further cluster talent and compute in a small number of dedicated hubs. The work done in these hubs should be open and transparent where responsible and highly secure where necessary. The private sector should be involved to help commercialise foundational research and to spread investment risk by providing additional funding. Finally, the benefits brought about by the institute should be shared fairly among participants, which could - at a later stage - also involve like-minded non-EU countries.

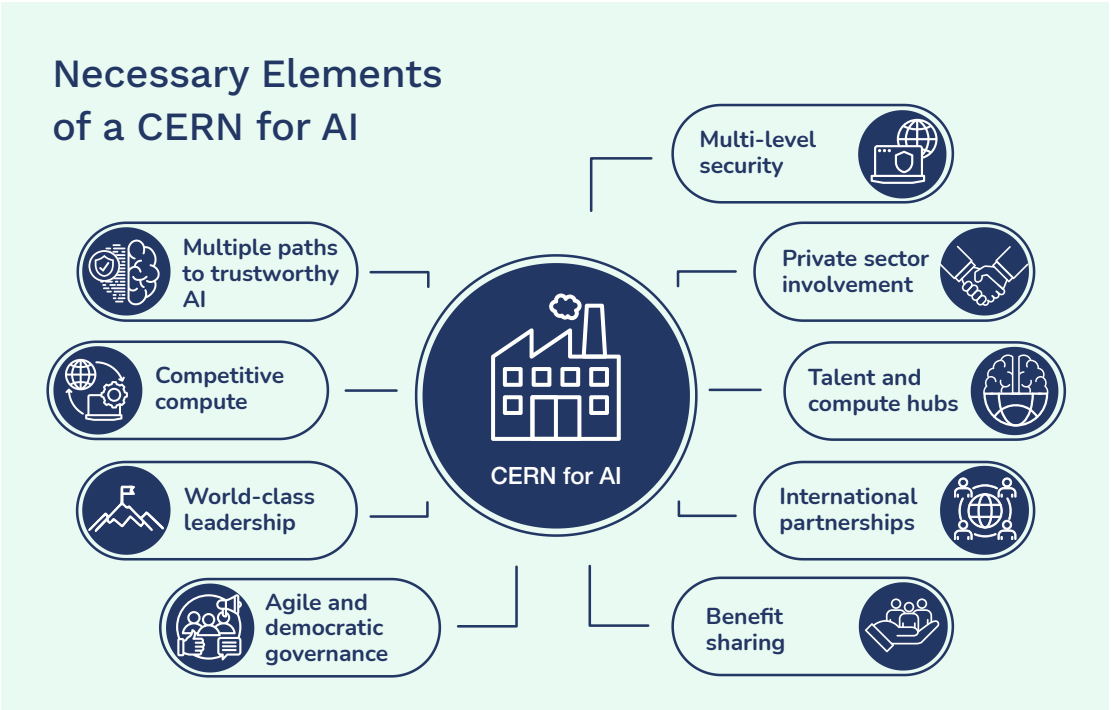

Figure 4 (right): Nine necessary components of a CERN for AI

Pursue multiple research bets with trustworthiness as a north star

A CERN for AI would need to adopt a portfolio approach to its fundamental research. The EU shouldn't try to beat American AI players at the current AI development paradigm - these companies have more experience, can draw on years of confidential algorithmic progress and will continue to have access to superior infrastructure. Instead, the EU should try to topple the game board by pursuing a portfolio of under-resourced, innovative research paradigms, with trustworthy AI as a north star.

Trustworthy AI is currently an unsolved scientific problem. That means any sufficiently ambitious approach has a non-negligible risk of failure. Most private companies aren't positioned to take this risk, and instead are mostly copying each other's strategies. The result is that their approaches are strongly correlated along the path of least resistance, which, unfortunately, is unlikely to lead to truly safe and trustworthy AI. By exploring several neglected paradigms at once, the EU can provide a counterpoint to the herd mentality of private companies, giving it a bigger chance of inventing trustworthy AI.

Such a portfolio approach could be implemented by 3-10 different work programmes. Leadership would be in charge of allocating computational resources among these different programs. It is too soon to tell which research agendas would fit this purpose, but it is possible to sketch out rough directions. The text box below provides a couple of examples of promising work programmes that CERN for AI could feasibly lead on.

First of all, current approaches in advanced AI suffer from poor reliability and a lack of rigorous reasoning. Neurosymbolic approaches, such as Google DeepMind's recent AlphaGeometry 2, could remedy these flaws. Another promising direction, that recently received 59 million pounds in funding from ARIA (Advanced Research and Invention Agency), aims to create safe AI systems by design in a provenly safe manner (i.e. making it mathematically impossible that a system behaves outside of specified constraints). Turing Award winner Yoshia Bengio has recently joined this effort as Scientific Director. A third strand of research could focus on creating 'bounded' systems that mimic human reasoning to ensure that AI systems do not behave in unpredictable ways.

A fourth research avenue could focus on mechanistic interpretability, a set of techniques that aims to uncover and characterise hidden patterns inside AI systems, like a 'digital neuroscience'. Manipulation of these patterns can change the AI model's behaviour and be used to steer answers in, for instance, more honest directions. Anthropic recently released a toy application of this technique by creating 'Golden-Gate Claude' an instance of Claude 3 Sonnet that would try to steer any conversation towards the topic of the Golden Gate Bridge. Mechanistic interpretability has recently seen a number of large breakthroughs, but further progress seems to require large amounts of compute. This makes it a potentially perfect fit for a CERN for AI.

A CERN for AI could still train large, multimodal models like the ones currently on the market. Such models can very well serve as elements of more elaborate system designs. Furthermore, training large models will yield economic value - especially if they can be finetuned and personalised for different European audiences - and will enable researchers to build experience with large-scale projects that require tackling difficult hardware challenges. It is crucial though, that the institute doesn't lose track of the overarching goal, which should always be to invent truly trustworthy AI. CERN for AI has a unique opportunity to diversify the advanced AI landscape. With such a transformative technology, society shouldn't put all its eggs in the same, corporate basket.

Ensure access to frontier computational infrastructure

In order for a CERN for Trustworthy AI to bear fruit, it requires scale - a lot of scale. A CERN for AI can put the EU back on the map, but only if the EU decides to invest at a much bigger scale. This applies predominantly to AI-infrastructure. The EU has a large amount of high-quality data at its disposal, but currently lacks the computational infrastructure to turn this resource into valuable AI products. To be competitive in infrastructure with the leading private companies by 2026[4], a CERN for AI would likely need to acquire some 200,000 NVIDIA GB200 superchips[5], all located in a maximum of 5 geographically separated campuses that are linked by high-bandwidth connections. This is an order of magnitude more ambitious than the compute investments from the EuroHPC AI Factories program. Estimated costs for such an effort amount to roughly 30-35 billion euros, including operational costs, personnel costs and energy costs over a three-year period[6]. While this is a large sum of money, it is comparable to existing programmes such as the EU Chips Act. In fact, large, public AI infrastructure investments form a logical continuation of the Act: Europe not

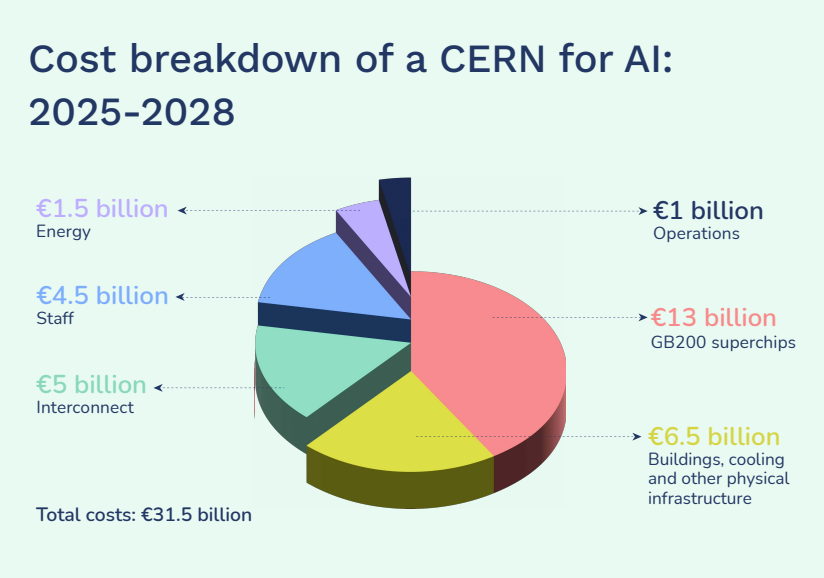

Figure 5: Breakdown of the costs of a CERN for AI.

[4] OpenAI and xAI are scheduled to have access to 100,000 GB200 campuses by 2025, and OpenAI and Microsoft have plans for much bigger buildouts in the years after.

[5] Or equivalent hardware from other providers. Note that Google DeepMind has ordered at least 400,000 of these GB200 superchips, in addition to the millions of TPUs they design in-house.

[6] Cost estimates are based on the following assumptions. A single NVIDIA GB200 server is estimated to have a price tag of around 65,000 EUR; 2) Within-rack interconnect is estimated to account for an additional 28% markup. Between-rack interconnect is estimated at another 15% markup. The physical buildings, cooling systems, on-site power machinery and other physical overhead is estimated at 14 million per MW of critical IT equipment for a total of 470 MW. This yields total capital expenditures of around 24.5 billion EUR. Energy costs are assumed to equal 0,07 EUR/kWh over a three-year period, assuming 80% utilization and a PUE-ratio of 1,25. Personnel costs are estimated at an average of 500,000 EUR per year, for a total of 3,000 staff members. Other operational costs such as data center maintenance and governance overhead are estimated at a total of 1 billion EUR over 3 years. This yields a sum-total of 31.5 billion EUR over a three-year timespan.

only needs a domestic chip manufacturing industry, it also needs a domestic AI industry that builds on top of those chips. It is true that a CERN for AI could also simply rent AI chips from foreign cloud providers, but this approach could be risky as it would introduce the same dependency and security risks seen in the broader semiconductor supply chain.

A CERN for AI would require state-of-the-art computational infrastructure for three reasons:

1. **Compute enables scaling.** Extensive research shows that AI models perform better with increases in size (e.g. their number of parameters) and the number of data points they are trained on. Both of these require compute. Scaling laws have been found for many different learning architectures and modalities. Whatever research direction proves most promising, it is thus a safe bet that it will require considerable amounts of computational resources to scale all the way to competitive models.

2. **Compute attracts talent.** If a CERN for AI wants to attract leading talent in advanced AI, it must be able to promise researchers access to large amounts of computational resources. The outpour of academics joining private companies has been largely driven by a desire to test ideas at larger scales that cannot be accommodated by academic institutions (funding constraints in academia have led to a so-called 'compute divide'). Indeed, many AI companies market themselves as an attractive employer by pointing to their superior compute resources.

3. **A portfolio approach requires compute.** Finally, in order to pursue multiple, parallel research bets towards trustworthy AI the EU needs access to more computational resources than making a single research bet. Although it would be possible to pursue different research bets in series, this would slow down progress to such an extent that it seems unlikely the EU would be able to catch up.

Highly distributed efforts don't add up

Given competing interests of Member States and difficulty in securing sufficient clean electricity, it will be tempting to invest in computational resources in a highly distributed fashion - i.e. spread out over all participating countries. However, this would likely entail an unacceptably inefficient use of public resources. Because AI chips need to 'talk to each other' at high speeds during large, multi-chip training runs, separating infrastructure this far degrades training efficiency to a restrictive degree.

The EU needs to be pragmatic and invest in 1 to 5 campuses[7], each of which houses several data centres that are seamlessly interconnected by state-of-the-art optic network equipment.

[7] At 40,000 GB200s per campus, each campus would still be able to train systems with compute budgets matching AI models currently in development by the leading private players.

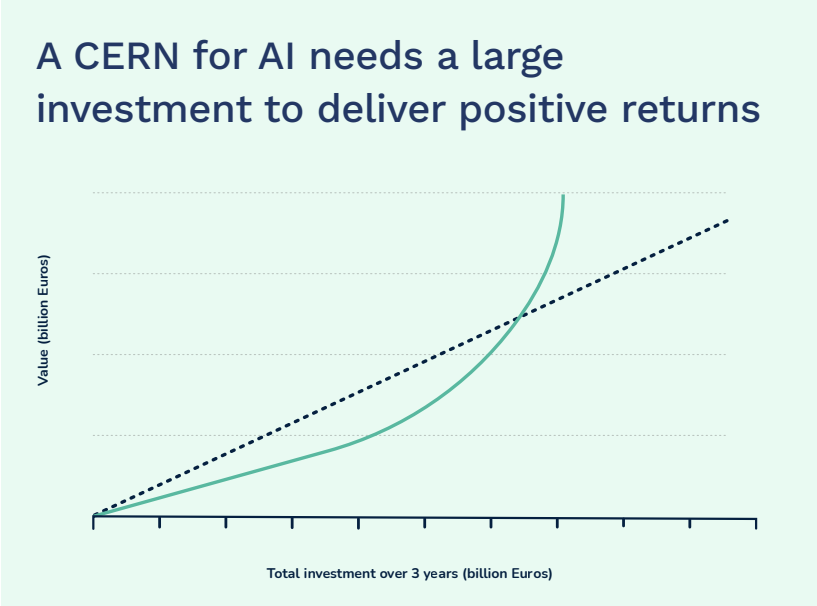

Figure 6: A CERN for AI needs to be big to generate a positive return on investment. (Numerical values for illustrative purposes only).

Some will react to this call for a large, relatively centralised investment with scepticism. However, in reality, large-scale investment is the conservative option. Efforts that fail to bring enough funds to the table, or which distribute computational resources over multiple language-specific training runs have a much lower chance of yielding a positive return on investment. After all, such efforts will most likely result in inferior models and products. In a winner-takes-most market, suboptimal solutions may see negligible adoption and thus fail to generate barely any value. By investing with ambition, the EU has a much bigger chance to invent trustworthy AI. And, even if it turns out that this ship has already sailed, most infrastructure investments can still be redeemed by renting out the GPU clusters to private companies and academics who are still starving for compute.

Appoint strong, entrepreneurial leadership

Having ambitious goals and actually realising them are two different things. Strong and entrepreneurial leadership will be crucial in actualizing CERN for AI's goals and operationalizing its vision. CERN for AI cannot be bogged down in bureaucratic processes while the wider AI industry is operating at breakneck speed. Once CERN for AI is decided upon, picking leadership that can hit the ground running should be the first priority.

There are three key leadership roles that CERN for AI will need to fill:

1. Research Leadership
2. Infrastructure Leadership
3. Political Leadership

Research leadership is needed to attract talent

In order to attract top talent, CERN for AI needs to pick respected and accomplished scientists who can inspire and encourage researchers to leave industry and move to Europe to work for CERN for AI. This will involve selecting a Chief Scientist and a team of Programme Directors.

This is similar to how the UK and US AI Safety Institutes (AISIs) hired prominent AI researchers like Geoffrey Irving and Paul Christiano to successfully attract top

talent. Another example is the UK's Advanced Research and Invention Agency (ARIA) appointing Yoshua Bengio (one of the godfathers of AI) as scientific director. These high-profile scientific hires were instrumental in these institutions successfully attracting leading technical talent. CERN for AI would have to mirror this strategy in order to stay relevant in the intensely competitive AI talent pool.

Infrastructure leadership is needed to expedite construction

At its heart, CERN for AI is an infrastructure project. All of the research and funding will be for naught if the computational infrastructure necessary to train and run the AI models isn't built expeditiously. Therefore, CERN for AI will need to bring upon leadership with experience building out large-scale infrastructure projects quickly and effectively.

The head of infrastructure will also have to be able to work closely with industry, given how much technical expertise related to chips and data centres is currently accessible only in the private sector. In fact, because of the historic lack of investment by the public sector, the specific expertise needed in running and building such clusters mean that this hire will likely have to come directly from one of the existing large technology companies - there is a relatively small pool of qualified candidates.

Political leadership is needed to minimise delays

Research and infrastructure leadership, while important, will be for nothing without political leadership. At the end of the day, many challenges that a CERN for AI could face can only be resolved at the political level. Picking a CEO/chair who can navigate the political landscape in order to help CERN for AI achieve its goals will be crucial. Only if leadership has sufficient political buy-in can a CERN for AI achieve its goals of reigniting the European economic engine, strengthening European security and safeguarding trustworthy AI.

CERN for AI's governance structure needs to be ironed out

Of course, outside of these leadership positions there are many unanswered questions on the governance of a CERN for AI. How, exactly, will leadership be appointed? How will national governments be represented? What kind of formal body should CERN for AI become? Such questions are outside the scope of this report but will be addressed in a forthcoming follow-up publication.

→ Create dedicated talent and compute hubs

While compute at scale is necessary for CERN for AI to achieve an impact, no single AI product will get built without the other key input to AI systems: talent. There are many calls for a CERN for AI to adopt a decentralised structure. While there are several benefits to this approach, intense AI talent shortages means it falls short in practice. In von der Leyen's political guidelines, her big AI proposal for the EU was to "pool all of our resources, similar to the approach taken with CERN." Alongside funding, talent is another resource that needs to be pooled and centralised to stay competitive.

An AI talent shortage makes centralization necessary

The supply of AI talent is painfully limited. Most of it is concentrated within AI and Big Tech companies in a few key cities in the US and Western Europe. A recent study indicates that 55% of top AI talent is located in just the US and UK. To make things worse, the demand for AI talent is extremely high (and continues to grow) both within industry, government departments and AI Safety Institutes (AISIs). Centralization is a natural solution to this problem. By centralising and clustering talent in one location, insights can be shared across multiple competing demand sources. There's a reason tech startups cluster around Silicon Valley and California: they benefit massively from the existing talent density and agglomeration.

A CERN for AI needs to adopt the same approach when it comes to its talent pool. While a decentralised approach could benefit from more diversity of opinion, broader accessibility and more appetite for experimentation, the reality of the situation when it comes to AI's steep competitiveness for talent necessitates a more centralised approach. The bottom line: if CERN for AI cannot bring in the requisite talent to compete, the project will fail.

However, centralization of the talent pool doesn't mean centralization of the benefits that come from a CERN for AI. For one, all of the open research carried out will be able to be distributed across Europe. Datasets generated by a CERN for AI will be available to academics in research communities across the continent to help empower their work. And closed research will also benefit the wider-Europe, with applications able to be built outside of the core talent hub through remote access (APIs). The centralised hub can be seen as a creator of technological infrastructure that Europe can build on. And, of course, all members of CERN for AI will be part of its benefit-sharing program.

Some locations are better than others

An AI talent hub would benefit from good transport connections to neighbouring countries, existing talent density, a metropolitan environment where international researchers can integrate seamlessly, and a surplus of amenities to help attract talent.

Compute hubs, on the other hand, need large amounts of space, clean-firm power sources (like hydro- or nuclear power), powerful internet connections, and good infrastructure and logistics networks. Luckily, the talent and compute hubs can be separated. To use the US as an example, most talent is densely located in California, with a high quality of life, whereas the data centres are largely located in Arizona, a comparatively less populated and attractive location for talent.

Advanced AI no longer relies on cheap labelling

In the early days of scaling up large AI models, AI companies extensively relied upon cheap overseas labour to label training data sets. However, as AI models have improved, they can now automatically label data sets and generate synthetic data, improving off their own outputs. This has seen the talent used in training pipelines shift towards highly-educated workers who can assess AI models on harder, more technical tasks, like checking code outputs for bugs or mistakes. In this domain, Europe's highly-educated workforce gives it a comparative advantage.

Invest in multi-level security

The research done by CERN for AI is likely to generate significant economic and security value. This makes it a lucrative target for espionage and theft. CERN for AI should employ a tiered security and access structure to balance open and accessible research with keeping security-related work out of the hands of malicious actors. This structure will also guarantee that CERN for AI models are safe from tampering, therefore making them more trustworthy.

The US think tank RAND recently released a report detailing the different security levels private AI companies should be adhering to to secure their AI models against various threat levels. They classify threat levels by measuring **operational capacity (OC)** for attacks from OC1 to OC5, using rough estimates of a threat actors' financial resources:

Threat Actor	Operational Capacity
OC1	$1,000
OC2	$10,000
OC3	$1 million
OC4	$10 million
OC5	$1 billion

They also classify **security levels (SLs)** and which threat actors they can protect against.

SL1	A system that can likely thwart amateur attempts (OC1)
SL2	A system that can likely thwart most professional opportunistic efforts by attackers that execute moderate-effort or non-targeted attacks (OC2)
SL3	A system that can likely thwart cybercrime syndicates or insider threats (OC3)
SL4	A system that can likely thwart most standard operations by leading cyber-capable institutions (OC4)
SL5	A system that could plausibly be claimed to thwart most top-priority operations by the top cyber-capable institutions (OC5)

CERN for AI should employ a **two-tiered approach** to security. The lower tier should employ SL1-2 level security, which would apply to the bulk of the research that is sharable and open. The upper tier should employ SL3-4 level security, with research being secured from adverse actors.

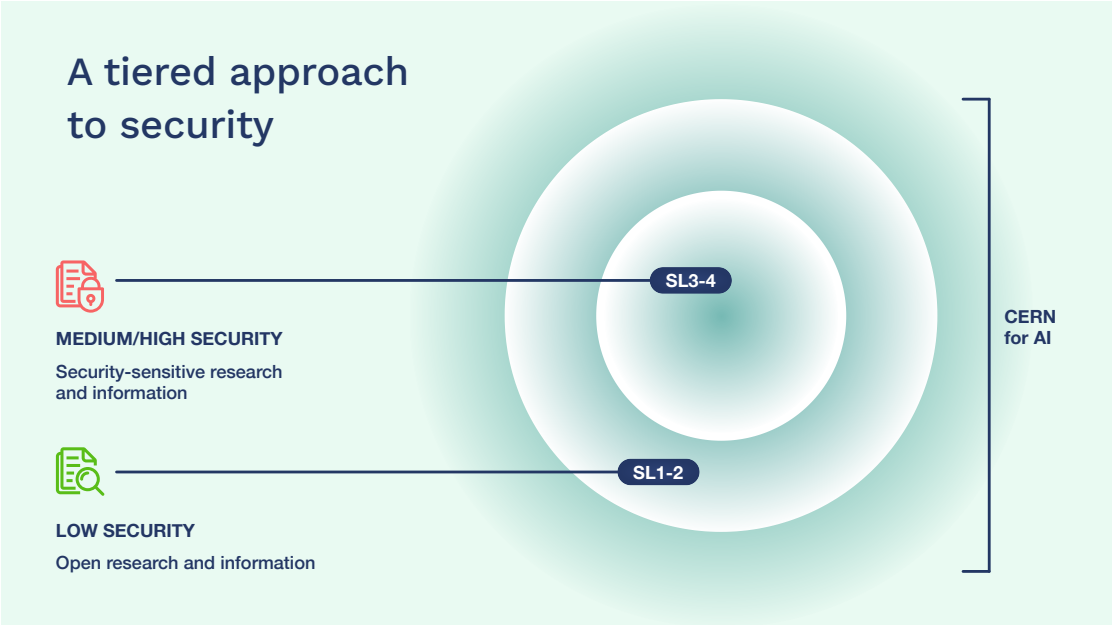

Figure 7 (right): A CERN for AI needs a tiered approach to security.

Upper tier security

Some of the research or programs that may fall in the upper tier could be:

1. AI models with any sufficiently advanced capability that surpass internally developed risk-thresholds. These would be similar to Anthropic's ASL3, Google DeepMind's Critical Capability Level 1 and OpenAI's "High-Risk" threshold, as well as other thresholds set to be decided at the Paris AI Action Summit.
2. CERN for AI's RAID unit (Risk Assessment, Information sharing, and Disclosure), to assess and analyse various AI-related risks posed by models and external actors, as well as act as an independent body that can be trusted by governments and industry.
3. Certain defensive technology programs (i.e AI-automated cybersecurity, disinformation flagging and intelligence gathering) that may involve sensitive information.

RAID (Risk Assessment, Information sharing & Disclosure Unit)

In order to properly mitigate external threats, CERN for AI should create a RAID (Risk Assessment, Information sharing and Disclosure) unit. This coordinator, housed in the portion of CERN for AI with heightened security, should take in information about AI threats from external actors and frontier models and use cutting-edge analysis and evaluation techniques to assess risks and report them to the relevant national stakeholders. This unit should facilitate the coordinated disclosure of dual-use capabilities present in AI models.

The RAID unit will need to:

- Develop in-house expertise and draw from international AI talent in order to properly assess AI-domain risks and interpret external incident reports, similar to the UK AI Safety Institute.
- Remain independent from regulatory agencies, to remove disincentives for industry actors to report incidents to them.
- Act as a coordinator, maintaining strong relationships with both industry, third-party risk-assessors and evaluators, and relevant government bodies.
- Triage and pass on reports on sectoral-specific risk to relevant actors (e.g. forward findings on an AI-enabled cybersecurity incident to the European Union Agency for Cybersecurity)
- Maintain ownership over AI-specific risks, like deceptive or autonomous behaviour.

CERN for AI's upper security level can build upon the EU's existing strong cybersecurity practices, the EU's cybersecurity act and contribute to two key components of the EU's cybersecurity strategy by building:

- resilience, technological sovereignty and leadership (by securing EU strategic autonomy);
- operational capacity to prevent, deter and respond (with defensive technologies and RAID);

Specific measures are necessary to secure the upper tier of CERN for AI

The level of security required in the upper tier is contingent on the resources spent on CERN for AI, the capabilities of AI systems being developed and the information being generated, analysed and distributed as a result.

If CERN for AI becomes a multi-billion-euro project, then it should expect OC3-4 level attacks. This calls for SL3 level requirements for securing model access as an immediate priority in order to ensure the resulting models and algorithmic secrets don't get stolen by adverse actors. Such security includes centralised and restricted management of weight storage, increased data centre security, with full-time security and inspections for unauthorised access or intrusion, and protocols and restrictions in place to decide which employees have access to model weights.

Beyond securing model weights, CERN for AI will also have to include SL3 level security for its network and non-weight sensitive assets, like algorithmic improvements as well as security assurance and testing, and threat detection and response, which could involve placing effective honeypots.[8] This is to ensure sensitive information isn't leaked or stolen by competitors or adverse actors. CERN for AI will also have to include SL4 level personnel security, such as occasional employee integrity testing and an advanced insider threat program, in order to ensure the security of sensitive information being shared within its information & risk assessment unit. CERN for AI's infrastructure will only be as secure as the individuals managing it.

Because of the sensitive nature of upper-tier research, any governments wanting to gain access to the research or information generated will also have to meet stringent cybersecurity thresholds.

As research continues at CERN for AI and it moves on to later phases, the economic value and power of the systems being developed will continue to increase. As a result, security levels may have to further increase in the upper tier to SL5, or a tertiary, even more secure, tier may have to be created, as more valuable research becomes a more lucrative target for sabotage and theft.

Lower tier security

All remaining research at CERN for AI would take place in the lower tier, with far more open access and collaboration. This could entail safety work like interpretability research, or applications of models for real-world use-cases, under a licensing system or API-access to closed models for productization of models. Private companies would be allowed partnered access to research and information within this tier.

→ Leverage public-private partnerships

A CERN for AI must collaborate with the private sector to succeed. The private sector brings the know-how to undertake such an ambitious project, with companies often having access to proprietary data, supporting infrastructure, and experienced staff. All of these can help CERN for AI hit the ground running.

Second, the private sector can bridge the gap between foundational research and value-providing applications. Companies pursue market needs and, therefore, strongly focus on product development and commercialization. Their participation makes it more likely that AI innovations are practically viable and market-ready. Historically, the private sector has been great at integrating several publicly developed foundational technologies, and turning them into groundbreaking products (perhaps the best-known example is Apple bringing together innovations like GPS, the internet, and touch screens into the smartphone).

[8] Decoy systems used to detect and analyse unauthorised access attempts or cyberattacks.

Third, the private sector can provide valuable funding in a tight moment for public budgets. Considering that Horizon Europe's budget is over €95 billion for seven years, a €30-35 billion three-year investment is difficult but possible—with the right financing strategies. However, simply reallocating a part of the EU budget approximately equal to the yearly cost of the entire European public administration to a CERN for AI is unrealistic. Giving the EU more own resources has also been a challenging process that will almost certainly fail to bring €30-35 billion to the table in the short term. Common borrowing among EU member states emerges as an option. The EU has already taken a substantially larger loan: during the COVID-19 pandemic, EU leaders agreed on a €700+ billion common borrowing. Nevertheless, common borrowing is not free and comes with financial risks to taxpayers. The EU must issue solid repayment plans concurrently with a CERN for AI funding proposal. This requires that the institute is designed to pay for itself. Still, a risk persists. An alternative to common borrowing is the Commission, EU member states, and companies pooling their resources. A large portion of the moved capital could come from the private sector in line with large-scale tech hardware investments like the EU and US CHIPS acts.

Nonetheless, public-private partnerships pose various risks and challenges that governments must address. The differing fundamental incentives between private and public actors, between profit and the 'public good', can create conflicts of interest. Authorities must ensure that participating companies have the incentives to meaningfully contribute to socially beneficial project goals. More concretely, certain participating companies may attempt to sacrifice trustworthiness for speed of AI development. CERN for AI's leadership should put in place guardrails to prevent this, as trustworthiness is a precondition for widespread technology adoption.

→ Take an adaptive approach to open source

Whether frontier AI models (at CERN for AI or otherwise) should be open-sourced remains an ongoing debate within the research community. While open-source models provide benefits along the axes of transparency, adaptability and distribution, a significant number of experts worry that next-generation open-source models may be seriously misused by malicious actors. The US' National Telecommunications and Information Administration (NTIA) recently released its report that found that the evidence base does not yet support substantial restrictions on open weight models. Current models are probably safe to stay open.

However, they also find that the evidence base does not support never applying restrictions. Some future, more capable models may have to be closed source. Future open-source restrictions will be contingent on collecting more evidence about potential threat models posed by AI systems. A CERN for AI could help build the evidence base that informs legislators when open-source restrictions are necessary.

Whether open-source models continue to grow more capable and ultimately present dual-use capabilities also significantly affects CERN for AI's internal strategy. It is worth exploring these effects under three different scenarios.

SCENARIO 1: Frontier open-source systems are released by the private sector and used by CERN for AI

If open source AI models successfully compete with closed models at the very frontier, it makes less sense for CERN for AI to exclusively build its own frontier AI systems from scratch. After all, this would amount to lots of double work without much benefit. Instead, more emphasis should be placed on building upon these open systems as a starting-block and on several targeted research streams that stem off from this foundational research. In this scenario, CERN for AI resources should be spent funding:

- Research streams focused on building out applications from existing open models
- Research focused on expanding the safety and capabilities of existing open models
- Novel and under-explored research directions and paradigms for safety and capabilities (which would also encompass training models from scratch)

One caveat is that while many of the concerns about strategic autonomy are resolved by open models, some security concerns may still remain. Namely:

- Is CERN for AI research on open models public, or is it simply used as a starting point for closed research? If this research is public, it could end up aiding adverse actors.
- What licence arrangement will CERN for AI secure with open model creators? META has already restricted the Llama 3.1 licence in the EU - could this be exacerbated in the future? The negotiation around this licence would be a key detail.

SCENARIO 2: Open-source systems pose dual-use capabilities, limiting their distribution and role in CERN for AI

If society gains evidence that the next generation of models possess dual-use capabilities i.e the ability to exacerbate CBRN-threats (e.g. by helping to design bioweapons), open model releases should be restricted.

Under this scenario, CERN for AI will have to use its resources and talent to develop the next generation of closed models itself, within the upper tier of its security structure. Existing research on smaller models, novel research directions and applications of models can continue within CERN for AI's lower tier.

SCENARIO 3: Hybrid developments

These two scenarios are not mutually exclusive. A situation could easily develop where CERN for AI starts off working primarily off of open models and then pivots towards a more hybrid model, with some research open, but the most capable research with dual-use capabilities siloed off into the upper tier.

The Center for Strategic and International Studies conducted an extensive analysis on how open-source systems interact with security and defence priorities. They

found that open-source systems play a key role in current defence systems and preliminary research shows that open-source AI systems could offer similar benefits. However, they also acknowledge the possible risks open-source systems could pose, especially if they reach dual-use capabilities. Overall, they found there was a significant gap in risk-benefit assessments.

→ Create a robust benefit sharing system

For CERN for AI to succeed, the fruits of its labour need to be shared fairly among members. Benefit-sharing could be operationalized through a shareholder system, designed to benefit smaller countries while still encouraging greater investment by larger countries. However, there is also room for other, more novel mechanisms.

In order for a shareholder system to work, CERN for AI will have to monetise its value. CERN for AI's foundational research will already benefit the economy on a wider scale (e.g: productivity and efficiency gains in the public sector). However, CERN for AI will also carry out applied and application-driven research, which can be more directly monetised. These revenue streams can then either be reinvested into further research, or be used to lower individual taxpayer bills.

CERN for AI can choose to licence open releases and applications, and offer access to closed products via paid access to APIs. Of course, there are countless other ways to monetize its foundational work and leadership should favour an adaptive approach.

→ Remain open to international partnerships

While CERN for AI would start off as a European project, it should be open to future, non-EU expansion. Horizon Europe is a great example of the benefits of such an approach, with Canada and New Zealand contributing to the budget, and increasing the program's impact. Making CERN for AI an expandable institution means balancing the incentives for new members to join (namely, benefit sharing), and the incentives for existing members to accept new additions.

One way to manage this could be by allowing easy access to CERN for AI's lower security tier for new members, but having more stringent rules in place for CERN's upper tier. Under such a construction, access to the upper tier would not only necessitate meeting stringent cybersecurity requirements, but could also require meeting requirements relating to an applying country's rule of law and freedoms.

What a CERN for AI would bring to the EU

SECTION SUMMARY

In the coming years, Europe will face tremendous challenges in economic competitiveness, security and responsible technology governance. The domestic development of trustworthy advanced AI systems could provide a lever to tackle all these challenges in one sweep. For the EU, however, this is only possible through a moonshot effort: a CERN for AI that substantially contributes to the EU economy, provides the foundation for resilience-enhancing technologies, and steers the trajectory of advanced AI in a trustworthy direction.

Advanced AI could make or break the European economy

Europe's economy missed out on multiple tech-driven booms over the past 30 years, a key factor in the bloc's lagging economic performance. The EU cannot afford to let advanced AI pass it by too. CERN for AI can enable the EU to become a key player in the advanced AI development, spurring the development of innovative products and restoring trust in European investments. Time to catch up is quickly running out. But the new Commission has a chance to unlock substantial economic benefits from European-made AI.

The EU economy has been underperforming while missing multiple tech revolutions

Against the backdrop of underinvestment, a fragmented digital single market, a shrinking workforce, and the continent's industrial slump, the EU needs to leverage new potential sources of growth. If not, the Union will be unable to afford policies targeting its ageing populations, a globally intensifying climate crisis, a growing shortage of affordable housing, persistently high energy costs, and supporting Ukraine. While Europe's economy is struggling, the US and China are pulling away. Diverging GDP growth rates have become evident in the last 20 years, coinciding with the development of critical and high-value technologies such as computer infrastructure and web-based enterprises. Europe has failed to extract sufficient value from these high-tech industries. Moreover, even when European researchers contribute to ground-breaking new technologies, like the Internet, the benefits largely accrued elsewhere. A CERN for AI could reverse both of these trends.

The EU cannot afford to miss the advanced AI boom

Advanced AI could add substantial value to EU economies. While advanced AI's exact economic value-add is unclear at this stage, the forecasted scale and growth so far are hard to overstate. PwC estimates the advanced AI market will grow to the current combined size of the ten largest EU economies by 2030. McKinsey estimates that generative AI - only a subset of advanced AI - will become a multi-trillion-euro market. Moreover, advanced AI systems are already an integral part of many workflows in other sectors. For example, AI tools permeate marketing, software engineering and customer operations. Soon they could become crucial in sectors such as healthcare, transportation and advanced manufacturing. With the rise of next generation, increasingly agentic, AI systems, the economic potential could be even larger.

AI driven productivity gains could help the EU reverse its current economic malaise, but only if the bloc secures a seat at the table. In a winner-takes-most market - currently dominated by American companies - a large percentage of advanced AI's value-add is likely to flow to non-EU players. Without a European alternative, the continent is susceptible to future price hikes that hit the entire economy. Worse still, the EU could lose access to foreign advanced AI models completely. In fact, Europe is already beginning to miss out on US AI models with Apple and Meta withholding products from the EU market. Losing access to a general-purpose technology like advanced AI in the 21st century would be like losing access to electricity in the 20th century. It would be a serious threat to the Union's economic prospects. The implications are clear: the EU needs to diversify its access to one of the most transformative technologies of our time.

A CERN for AI can boost growth, bringing significant spillover benefits

So far, Europe is stuck in making half-hearted efforts to develop advanced AI. If it wants to be a key player in the technology's creation, it needs a larger, more targeted approach.

With the right investment and the right design, a CERN for AI can deliver world-class advanced AI that is more trustworthy than foreign alternatives. It can overcome infrastructure and talent constraints that have been holding back the European advanced AI industry. By bridging the gap between foundational research and applications, it can generate valuable, commercial tools that benefit both the private and public sector. And by taking gambles on more trustworthy design alternatives that aren't being pursued by private companies, it can promote uptake and ensure that AI technology is used for good.

A CERN for AI could also spur significant spillover benefits. For example, CERN developed the World Wide Web, which accounts for a whopping 2,9% of global GDP. Under the right conditions, with generous funding and encouraged experimentation, we get bold, groundbreaking, innovative products.

Finally, a successful CERN for AI can promote investor confidence. Europeans hold idle savings equal to a third of the US economy. If, instead, these savings were invested in high-growth sectors, they could rejuvenate Europe's economy. Moreover, EU pension funds take at least a fourth of their funds abroad. For Europeans to start

investing in their own region, they have to be shown the potential upsides. Currently, citizens worry that, as Commission President von der Leyen put it, Europe is "slow, burdensome, and distant". An ambitious and successful moonshot project could prove to Europeans that co-investing with the public sector can be wise. A CERN can demonstrate that Europeans haven't lost the ability to undertake world-class foundational and applied research, and that they can also commercialise it.

Time is running out

This might be the EU's last chance to catch up to foreign advanced AI developers. In the US and China, large-scale public and private investments are flowing into advanced AI and the semiconductors their AI models are trained on. Without a large, dedicated European effort, these nations will solidify their leads. Catching up will become more and more expensive the longer the EU waits.

European development of advanced AI is a geopolitical and security priority

Alongside economic growth, European security has risen to the top of the political agenda for this decade. The war in Ukraine, persistent cyber threats by China, and a possible isolationist US administration all emphasise the EU's need to protect itself. Advanced AI creates a unique set of challenges for individual states' security and the Common Security and Defence Policy (CSDP), but also offers novel solutions. The scale and talent density required to build these solutions necessitates acting at a European scale. However, Europe can't simply append "AI'' to existing security policies. Europe needs its own set of policy efforts specifically focusing on the intersection between AI and European security.

Advanced AI poses or amplifies two major security risks:

1. A lack of European strategic autonomy; and
2. Exposure to threats from external, hostile actors.

Europe needs strategic autonomy over AI systems used in critical infrastructure

Several security risks arise from the concentration of technological power and knowledge within a handful of private technology companies primarily located outside the EU. Europe needs the ability to steer the direction of technology and have stronger ownership over systems used in critical infrastructure and security systems.

AI's integration into the broader economy also means its eventual integration into critical processes and infrastructure. This diffusion will bring with it a novel set of security challenges. Namely, if AI models are integrated into infrastructure systems like the electrical grid, questions will be rightly raised over the origin of these models and their trustworthiness. The same security concerns that arose around Huawei's involvement in critical infrastructure during 5G roll-outs across the West should be considered now with advanced AI. To address the need for maximally reliable AI systems in critical infrastructure, there will be an important market

for AI systems that contribute to strategic autonomy. Having a publicly-backed institution build these systems is a plausible solution, and CERN for AI is the only current proposal with the sufficient scale to succeed.

Advanced AI will exacerbate existing external threats and create new ones

The widespread adoption of AI systems will enable a wide range of new threats and exacerbate existing ones. These threats need to be met with increasingly technological responses and risk assessment. In fact, the Commission already recommends risk assessments on AI as it is "considered highly likely to present [one of] the most sensitive and immediate risks related to technology security and technology leakage".

Hostile foreign actors can use AI to accelerate and expand their geopolitical influence, contributing to digital and hybrid threats. We have already seen early attempts, such as Russia allegedly using deepfakes of President Zelenskyy to convince Ukrainian soldiers to lay down their arms, or China using AI to expand its mass surveillance networks. Other threats governments are preparing for include AI-automated cyberwarfare or AI-mediated CBRN (chemical, biological, radiological and nuclear) threats.

While technology accelerates these threats, it can also be used to defend against them. AI systems could improve cybersecurity, flag and delete disinformation and enhance intelligence on terrorist threats. These defensive systems will only arise if we take the active choice to build them. To build this defensive technology, Europe needs a high density of top technical talent, as well as foundational research in AI systems, strong cybersecurity practices, and state-of-the-art infrastructure. A well-resourced and well-led CERN for AI uniquely positions itself to offer this.

There are many historical examples of the state driving the direction of technology and innovation. DARPA (the Defense Advanced Research Projects Agency) famously pushed forward key technologies like the internet and GPS, alongside CERN. In the UK, the Advanced Research and Invention Agency (ARIA) is mirroring this strategy. The success of DARPA led to the creation of ARPA-H (for health innovation) and ARPA-I (for infrastructure). CERN for AI should learn from these projects and pursue a similar vision, making targeted bets on high-risk, high-reward defensive technology projects that can help boost European economic growth and security.

Beyond bolstering security, CERN for AI can re-establish Europe as a thought-leader on the international AI stage. While the EU has shown itself to be forward-thinking when it comes to regulation, technology development and innovation has been lagging behind. This reinforces a damaging image of a Europe that seeks to govern a technology it has no hand in creating and therefore no understanding of.

Creating an institution like CERN for AI can shift this narrative to one where Europe is an active player in the room. Building CERN for AI can promote a perception of the EU as a major player that is participating in the building of the technology itself rather than a minor actor that has to rely solely on regulations to have impact. As AI improves, it will increasingly become a lever and a wedge that influences geopolitical discussions and agreements. Acting now can secure Europe's influence for decades to come.

Trustworthy AI requires democratic oversight

Private companies are failing to develop trustworthy AI solutions

As laid out in the EU's ethics guidelines for trustworthy AI, advanced AI can only be trustworthy if systems are shaped by a diverse set of stakeholders, if they are safe, robust and secure and if providers can be held accountable. Today, private companies are failing to meet these demands.

The AI market suffers from severe concentration of power

The advanced AI market is becoming more and more concentrated, with a handful of private companies dominating the field. Winner-takes-most dynamics are likely to increase this level of concentration even further. This translates to a virtuous cycle for incumbents: increasing returns from scale leading to increasing investment in scale leading to increasing returns from scale. Market consolidation - embodied by the recent 'acquihires' of Inflection AI, Adept AI and Character AI - implies that a set of 1-5 leading AI companies and their cloud providers (i.e.. Google DeepMind, OpenAI/Microsoft, Anthropic/Amazon, Meta and/or xAI) may decide over the trajectory of a technology that could shape the future of humanity unlike any other. Worse still, it seems that faulty internal governance structures, like toothless non-profit boards, are failing to prevent CEO's of these companies from asserting unilateral control over strategic decisions these companies face. Barring regulation, a handful of tech CEOs are currently given carte blanche to shape the future of AI, and, possibly, the future of society.

Frontier models inherently lack safety, reliability and security

The state of advanced AI can increasingly be characterised as 'a race'. Companies are racing to beat each other to market with new types of AI models. Simultaneously, American policymakers are enabling these companies, because they perceive themselves to be in a race with China to develop this transformative technology. While this race may be good for competition, it has also caused providers to cut corners on safety. In an effort to 'make Google dance', Microsoft released an early version of GPT-4 now widely known as Sydney Bing. Beyond its intended function of searching the internet and providing useful answers, it also tried to gaslight and deceive users to divorce their partners. Google DeepMind later released a model over-optimized for diversity that generated images of black people in Nazi uniforms, only to release a new model a couple of months later that advised users to 'eat a minimum of one rock per day'. While these safety failures are still regarded as suitable topics for Late Night monologues, accidents like these are already causing serious harms. Accidents involving next-generation systems could cause larger-scale, more severe and more widespread harms.

There are no signs that leading AI companies are replacing speed for care. OpenAI structurally failed to provide their Superalignment team with publicly promised compute resources for safety work because capability work was deemed more urgent. The Superalignment team has now been disbanded after an exodus of safety researchers. Meanwhile, an individual hacker stole internal secrets from OpenAI. In their latest Frontier Safety Framework, Google DeepMind explicitly states that their security is nowhere near the level sufficient to prevent state

actors from stealing model weights and key algorithmic secrets. With Google widely believed to have the best information security of all the leading AI companies, this shows just how inadequate the state of cybersecurity is in the AI market. And without proper cybersecurity, any safety mechanisms may easily be compromised.

Meanwhile, the players with the best chance of catching up, aren't doing much better. Meta and Mistral both saw their frontier models accidentally leak, and Meta and xAI have both made it increasingly difficult for users to opt out of the default setting that enables the companies to train AI models on their user data (e.g. tweets of facebook posts).

Leading developers are dropping the ball on transparency

Leading AI closed-source companies (i.e. OpenAI, Anthropic and Google DeepMind) are also becoming less transparent. Their decision not to release frontier model weights publicly is understandable, as they are valuable trade secrets that are resource-intensive to produce. However, opaqueness in other areas has resulted in an erosion of trust. For one, users are kept in the dark on what kind of data these models are trained on. When asked questions about training data, executives often resort to defensive and vague language (possibly because of the ongoing law-suits they face from several content-creators for breaching copyright). Meanwhile, safety reporting is increasingly pushed back to well after models are released. This lack of transparency also locks out independent academics and researchers.

Transparency is not only missing on a model-level but also at a wider operational and organisational level. OpenAI was found to be actively trying to silence ex-employees from critiquing the company with shady legal constructions. Employees were asked to sign non-disparagement agreements covered by other non-disclosure agreements, meaning they couldn't mention the fact that they weren't allowed to critique their previous employer. OpenAI even went so far as linking these non-disparagement agreements to employees' vested equity, effectively blackmailing ex-employees with large sums of promised equity (sometimes amounting to 90% of their net worth). While OpenAI leadership has since promised to remove these conditions and has said they were unaware of these clauses, new information on additional, probably-illegal whistleblowing clauses cast serious doubts on their intentions. All in all, lack of transparency and cover-up attempts are rapidly causing the general public to lose trust in leading AI companies, which was already lacking.

It shouldn't come as a surprise that private companies are failing to steer the trajectory of advanced AI in a more trustworthy direction. While governments are able to pursue something akin to the "common good" (although, granted, they often fail at this), leading AI companies are mostly optimising for a combination of shareholder value and prestige. This attitude is well summarised by the Silicon Valley phrase 'move fast and break things'. The solution to this mismatch is clear: Advanced AI development needs more meaningful democratic oversight. By increasing the level of public scrutiny, society can ensure that AI systems serve the common good.

European regulation won't ensure that AI is trustworthy

The EU's digital and data laws are a powerful set of tools that aim to compel actors to align with European values and prevent systemic harms. However the EU cannot bet on regulation alone.

The potential harms of unsafe AI systems are global in nature, so not unique to any one legal jurisdiction. If a developer in a non-EU country decides to ignore the EU market (if, for instance, EU regulatory compliance costs are deemed too high), the same product may still be misused to perform cyberattacks on European targets. With Apple and Meta choosing not to release their most recent AI products in the EU, it looks increasingly unwise to put all one's hope on the so-called Brussels Effect.

In addition, the pace of AI progress indicates that regulation could be too slow to properly steer developments. Drafting of the EU AI Act began in 2020 but some of the Act's requirements for certain high risk systems will only take effect in 2027. This seven-year lag between conception and full implementation is unacceptably long, given the trajectory of the AI market. Seven years ago, general-purpose AI systems didn't even exist yet. Even if AI Act revisions would follow a sped-up timeline, bureaucratic processes could likely prevent the EU from making the necessary changes in time.

While the rule of law is vital to ensure trustworthy AI, it must be complemented by effective stimulation. By participating in state-of-the-art AI development itself, the EU would gain a hand on the steering wheel, in addition to a backseat from which to call directions.

A CERN for AI would put democratic values at the heart of AI development

A CERN for AI would put meaningful democratic oversight at the heart of AI development, bringing three distinct advantages over the current private sector efforts:

1. **Diversity and inclusiveness.** Due to its pan-European nature, a CERN for AI can build on a wider, more diverse set of researchers during the development phase. Its organisational structure can further allow for a more inclusive process when it comes to strategic decision-making along the entire life-cycle of advanced AI systems. As AI isn't a value-neutral technology by default, such broad representation can increase trust in the resulting models and make sure the technology doesn't develop in a direction that is supported only by a small subset of society.

2. **Research-driven, not profit-driven.** Through its vast computational resources, a CERN for AI can pursue several, uncorrelated research bets that prioritise safety, sustainability, and reliability. Instead of trying to beat American AI companies at their own game, a CERN for AI could shift the technological trajectory in a more trustworthy direction.

3. **Transparency and accountability.** A CERN for AI can set a new standard in transparency and accountability, going above and beyond the requirements in the AI Act. If successful, this can speed up adoption of AI systems, particularly in the public sector, where legal concerns and privacy issues are currently holding back immense potential benefits. Setting the bar high could spark a so-called race to the top (in contrast to a race to the bottom), where other providers face economic pressure to raise their transparency ambitions as well.

Conclusion

Europe can get a seat at the table

The EU has the chance to found a fit-for-purpose CERN for AI. Any delays, however, could lead to the entrenchment of the leading positions of American and Chinese private companies.

CERN for AI is not just an ambitious vision, but a necessary step to secure Europe's economic future, safeguard its security and geopolitical standing, and steer the trajectory of AI development towards more trustworthy and ethically aligned systems.

With a targeted €30-35 billion investment, Europe can build the computational infrastructure at the necessary scale to compete in the AI economy and attract leading AI talent into creating this technology in Europe, for Europe. CERN for AI should follow public institutions like the UK and US AI Safety Institutes (AISIs) and UK Advanced Invention and Research Agency (ARIA) who have succeeded in talent acquisition by bringing on board high-profile researchers early on in the process.

In order to catch up on AI, Europe will need to take some targeted risks. By taking a diverse, portfolio research approach, CERN for AI will benefit from the diversity of the European AI and science community.

By picking select locations to centralise talent and infrastructure, Europe will be able to distribute the benefits of CERN for AI while reaping the benefits of agglomeration.

And with its multi-level security tiers, CERN for AI will be able to protect against novel threats and ensure that future AI systems are used for, rather than against, European citizens, while also ensuring that the research that can be open, stays open and accessible.

There are still many details to work out, but the foundations are in place for an institution that will not just stem a European decline, but spark a European technological renaissance. Yes, CERN for AI could be critical for addressing big questions around Europe's economy and security. But what's more exciting are the countless other problems it could solve. Now is the time for the EU's leadership to lead the way.

About the authors

DAAN JUIJN is a Compute Governance Researcher at ICFG. Previously, he worked at the Dutch Ministry of Economic Affairs, where he focused on the EU AI Act, AI scenario planning and the Dutch AI strategy.

BÁLINT PATAKI is a Senior Advanced AI Researcher at ICFG. Previously, he was an Accredited Parliamentary Assistant for MEP Maydell, where he focused on EU AI, biotechnology, and quantum policies.

ALEX PETROPOULOS is an Advanced AI Analyst at ICFG. Previously, he worked as a freelance journalist, writing and commentating about AI policy across TV, radio and print media.

MAX REDDEL is the Advanced AI Director at the ICFG. Previously, he was the Deep Uncertainty Lead at the Odyssean Institute, focusing on decision-making under uncertainty and societal resilience.

Acknowledgements

The authors would like to sincerely thank the following reviewers for their extensive feedback:

AARON MANIAM Fellow of Practice and Director, Digital Transformation Education, Blavatnik School of Government, University of Oxford

ROXANA RADU Associate Professor of Digital Technologies and Public Policy, Blavatnik School of Government, University of Oxford

ROBERT PRAAS Data Scientist, CEPS

CYNTHIA SCHARF Senior Fellow, Climate Interventions, ICFG

Image credits: Cover image generated with AI Shutterstock (reference to Giorgio de Chirico)

FOR MORE INFORMATION PLEASE CONTACT:

Max Reddel

Advanced AI Director

m.reddel@icfg.eu

International Center for Future Generations ICFG