SaferAI

The Case for European Investment in High-Risk, High-Reward AI Reliability Research

March, 2026

Table of contents:

Executive Summary

1. AI Reliability: from black boxes to Verifiable Guarantees
2. The Economic case: Unlocking Europe's Safety-Critical Industries
3. Strategic Benefits from Leading on AI Reliability
4. Europe’s window of opportunity (if it acts fast)
5. Market Failure: Why Private Investment Will Not Close the Gap
6. Existing European R&I Instruments Are Insufficient
7. Institutional design for “high-risk, high-reliability” research
8. Complementary Investments: Building a Reliable, Safe, and Secure AI Ecosystem.

Appendix A: estimating the size of “safety-critical” industries in the European vs American economies.
Appendix B: estimating the share of EU and US “industry champions” operating in safety-critical industries

Acknowledgments:

This paper was written by Chloé Touzet, Lily Stelling and Bruno Galizzi, from the SaferAI Policy team. We are thankful to several anonymous reviewers who commented on earlier versions of this paper. We particularly want to thank Daan Juijn, Senior Advanced AI Researcher at the Centre for Future Generations, for help with Part 6.

Executive Summary

Europe faces a structural challenge in AI adoption. An important share of its economy is concentrated in safety-critical industries, such as aerospace, rail, or energy, where current AI systems cannot be deployed because they lack the formal reliability guarantees that safety practices and certification require. Safety-critical sectors account for 10.2% of gross value added in the EU, compared to 8.3% in the United States, and European listed corporate champions are nearly twice as concentrated in these sectors as their American counterparts.

This paper argues that **Europe should seize a rare window of opportunity by investing in high-risk, high-reward research into verifiably reliable AI through an ARPA-like institution**, potentially as the core mission of the *Frontier AI Initiative*. This research would aim to produce *ex-ante* formal guarantees that AI systems will behave within specified bounds, enabling deployment in industries where the alternative to reliable AI is no AI deployment at all.

AI reliability research is a comparatively low-cost bet with outsized potential returns. The investment required is modest relative to the scaling-focused expenditures of frontier AI companies. An initial research program could begin at approximately €65 million per year, gradually ramping up to €1.8 billion per year, compared to the $100–200 billion that individual hyperscalers spend annually on compute infrastructure only.

The case for European leadership rests on three pillars. First, the **economic benefits** are substantial: reliability guarantees would unlock AI adoption across Europe's industrial champions. Second, the **strategic benefits** include reduced dependence on foreign AI providers in sovereignty-sensitive sectors, a powerful talent repatriation lever, and a leadership position in an emerging and potentially indispensable layer of the AI value chain. Third, a **market failure** in AI reliability research,driven by competitive dynamics, investor incentives, and sunk-cost lock-in among frontier providers, means that the private sector is structurally unlikely to supply this research at the necessary scale, creating an opening for public investment.

This paper outlines the economic and strategic rationale for European investment, diagnoses the market failure that justifies public intervention, identifies the institutional design features that would maximise the likelihood of success, and explains why existing research and innovation instruments are insufficient for this purpose.

1. AI Reliability: From Black Boxes to Verifiable Guarantees

Current AI systems are “black boxes.” We can observe their inputs and outputs, but cannot guarantee that they will consistently behave as intended or that they will not pose unacceptable risks. This opacity is not merely a theoretical concern, it is a binding constraint on deployment in industries where safety certification is mandatory.

Verifiably reliable AI refers to systems that come with *ex-ante* guarantees that the system will behave within specified bounds before deployment. This is sometimes called *safety-by-design.* It differs fundamentally from conventional “AI safety” approaches, such as red-teaming, reinforcement learning from human feedback (RLHF), or constitutional AI, which may improve average-case behaviour but cannot provide the formal guarantees required for deployment in aerospace, medical devices, or critical infrastructure. These *ex-post* solutions do not solve the problem of lacking *ex-ante* reliability[1].

The goal of AI reliability research is to build systems where scientifically rigorous claims of reliability and safety can be proven, in the same way that engineers can mathematically certify the structural safety of bridges or aircraft. This entails developing AI systems that can explain their reasoning in a form humans can verify, report well-calibrated uncertainty about their claims, and represent their beliefs honestly, avoiding deceptive, inconsistent, or unsupported outputs (see here for a more detailed research agenda).

Depending on the scale of investment, this research could yield two types of output: (i) a **minimum viable outcome** that delivers near-term benefits even if Europe is not the main producer of frontier foundation models, and (ii) a **more ambitious outcome** that could support genuinely safe-by-design frontier models developed in Europe.

The minimum viable outcome is a **European reliability layer that can be "added on"** to frontier models: methods, tools and reference systems that can evaluate, constrain, and verify the behaviour of frontier models (including those developed outside Europe), so that they can be deployed with credible evidence. In practice, this could include "reliability gatekeepers", i.e. models or systems that check the outputs of other models for inconsistencies, unsupported claims, unsafe plans, or violations of specifications.

[1] We acknowledge that earlier research agendas, including agent foundations research and previous interpretability work, pursued related ambitions and encountered significant tractability challenges. However, the agenda proposed here aims to produce *probabilistic bounds on failure rates,* for well-specified domains as a first step. It envisages a progression from narrow, domain-specific guarantees (e.g. for railway signalling) toward broader ones, delivering deployable intermediate results along the way.

The more ambitious outcome, if the reliability layer matures and Europe simultaneously invests in compute and secure infrastructure, is to **integrate these methods upstream into model training**, so that reliability, safety and security become intrinsic properties rather than patches. Such **homegrown safe-by-design AI models** would inherently reduce the risk of unreliable outputs.

The choice between these paths depends significantly on available compute resources. Importantly, these two endpoints are not all-or-nothing. A well-designed research programme can deliver valuable intermediate outputs[2].

2. The Economic Case: Unlocking Europe's Safety-Critical Industries

The reliability gap described above has direct economic consequences for Europe. Safety-critical industries represent a larger share of the European economy than of the American economy, and European corporate champions are structurally more concentrated in sectors that require formal reliability guarantees before AI can be deployed. This section sets out the economic case for public investment.

a. Unlocking AI adoption for Europe’s own industrial champions

The share of “safety-critical” industries in the economy is higher in Europe than in the United States: in 2023, it was 10.2% of gross value added in Europe, compared to 8.3% in the United States[3]. In addition, **Europe’s listed corporate champions are structurally more concentrated in industrial and safety-sensitive sectors** than those in the United States or China[4]. These sectors include rail, aerospace, or energy among others where *ex-ante* formal certification is required, and reliability and safety guarantees are needed before AI can be adopted. European champions in these sectors include, e.g., Airbus, Alstom, Philips, Thales, or Siemens.

[2] For example, starting with producing reliability guarantees in the form of probabilistic bounds on failure rates would already unlock many industrial applications, and is likely to be achieved more quickly than absolute reliability proofs.

[3] More precisely, that share is 10.17% of GVA in 2023 across the 21 EU countries for which data is available, compared to 8.34% in the United States that same year. These estimates are based on our own computation using Eurostat and Bureau of Economic Analysis data, see Appendix A for the detailed methodology. To calculate this, we operationalise “safety-critical” industries as follows: petroleum and gas refinement; chemicals manufacturing; pharmaceuticals manufacturing; aerospace, defence & other transport manufacturing; utilities; land transport (passenger and freight); air transport; and transport logistics.

[4] As an indicator, we compared the composition of the MSCI regional index in the EU , China and the US. The share of firms operating in critical sectors is 48% in Europe, compared to 26% in the United States and 20% in China. See Appendix B for a detailed methodology.

Key actors in these sectors have stated **they currently cannot adopt AI for safety-critical applications.** Standards governing automotive, industrial, or medical systems were designed assuming deterministic software. This is incompatible with today's AI models. The safety risks are too high to integrate models that are still unpredictable and unreliable in crucial processes. Major European company executives explicitly note these barriers:

- Alstom: "Currently, 80% of our resources go to non-safety-critical AI applications, which only make up 20% of our business". Alstom Chief AI & Data Scientist Nenad "Mijatovic: "the biggest opportunity lies in AI's ability to understand specific environments, like complex rail networks, and make explainable and safe decisions – grounded in expert-driven learning and quality data, with strong guardrails in place."
- Siemens CEO Roland Busch: "hallucination is not acceptable" in industrial deployments
- Airbus EVP Catherine Jestin: AI deployment must "comply with the regulatory constraints that characterise the aerospace industry," delaying large-scale deployment
- Philips CEO Roy Jakobs: "we need to unlock the application of AI, and we need to do so in a trustworthy manner", and Philips executive Jeff DiLullo: "the trust factor of the use of [AI] is actually quite nascent. It's the biggest barrier right now to larger scale deployment"

Thus, reliability guarantees would unlock whole categories of use cases in Europe. This could provide **a step change in productivity benefits for Europe**, as AI-enhanced industrial champions become more productive and competitive, supporting their exports and product quality.

Actors in safety-critical industries worldwide could benefit from these spillovers, but the case is strongest in Europe; because safety-critical sectors represent a larger share of European industrial champions, reliability guarantees are pressing and the potential economic value is proportionally greater[5].

[5] Not all of the value added in the safety-critical basket defined in Appendix A would be directly unlocked by reliability guarantees, the basket is an upper-bound proxy constrained by data availability. However, the argument is structural: Europe's economy is more concentrated in sectors where reliability is a binding regulatory constraint, and European corporate champions are disproportionately exposed to these sectors. Even if only a fraction of value added within each sector is directly affected, Europe benefits proportionally more from solving the reliability barrier than economies less concentrated in these industries.

b. Building an edge in the AI competition

Developing the research and innovation to produce *ex'ante* reliability guarantees of AI model behaviour in Europe would amount to securing leadership over a **part of the AI value chain that will become indispensable** once it is available:

- First, as argued above, powerful AI models developed elsewhere cannot be deployed in safety-critical applications without reliability guarantees; so the latter would be indispensable to unlock some of the potentially most productivity-enhancing usage of AI.
- Second, once it is possible to guarantee reliability *ex'ante*, it is likely that ***unreliable* models will become less attractive**, even in non safety-critical use cases, where unreliability is still costly. In other words, *guaranteed reliable* models could become more competitive.

If Europe gets a **headstart in reliability research and innovation**, it would build a **competitive edge** and increase its chances to capture much of the emerging market for reliability guarantees.

This research agenda would span multiple levels of the AI stack. At the foundation level, it would develop methods to verify and guarantee properties of general-purpose AI models (e.g. the approach pursued by the non profit LawZero). At the application level, it would develop guarantees for domain-specific systems that integrate AI into safety-critical workflows (e.g. formal verification for AI-enabled railway signalling or medical diagnostic tools).

Note that there may be some tradeoff between reliability and capability – for instance, systems with formal safety guarantees could be narrower in scope than unreliable frontier models. But for safety-critical applications, this tradeoff may be beside the point. The alternative to a reliable system is often no AI deployment at all. Hence, reliability guarantees would not constrain what is otherwise possible, but would enable what is currently impossible because the risks are too high.

c. Building the European brand and securing first-mover advantages

While some of the intellectual property could be protected through well-designed IP, the goal is not for Europe to permanently retain proprietary knowledge over AI reliability. Instead, we argue that investing in the research now and thus **getting a headstart could help i) build the European AI reliability brand and ii) secure a first-mover advantage**.

i) Building the European brand. Reliability could become a distinctive European brand, a good fit with the traditional "European quality" brand positioning (indeed, European manufacturing often competes on quality and reliability). The European offer could leverage European domain expertise in safety-critical sectors to build **domain-specific reliable AI packages**, using insights from trained practitioners, specialised datasets, and domain-specific safety specifications. Synergies with the EU AI Act could also be exploited, since ex ante reliability verification technology could be exported as **compliance solutions**. This positioning would also be a relevant **complement to the *ApplyAI* strategy**, by providing actionable, technical solutions to making Europe a leader in AI deployment.

ii) Securing first-mover advantages. History shows that early movers in tech innovation capture a lasting advantage. Deep-learning was considered high-risk, high-reward research before 2010. OpenAI, as a nonprofit, took moonshot bets with a team of excellent researchers between 2015-2018, to advance and invest in various promising research strands within the deep learning paradigm, including scaling self-supervised learning on the transformer architecture, among others. OpenAI and other early investors in this paradigm, captured enormous value as a result, and by the time the paradigm shift was undeniable, talent and infrastructure had concentrated in these institutions, creating first mover advantages that competitors now struggle to overcome. AI reliability occupies a similar position today, offering a **rare opportunity to lead rather than follow**. First movers' durable advantages include:

a) tacit knowledge accumulation: investment builds organisational capabilities, agglomeration effects, supply chain relationships, research infrastructure and, crucially, talent concentration. Investing early increases the chances of becoming an "AI reliability hub". This learning curve advantage persists even as competitors enter.

b) market trust accumulation: first movers with track records build trust that late entrants cannot quickly replicate. This links back to the idea of AI reliability as a 'European brand' explained above. Trust accumulation for first movers could generate brand recognition and loyalty to European technology as a default provider of AI reliability guarantees / guaranteed-safe AI models.

c) the potential to shape standards: first movers in AI reliability can shape emerging standards by influencing the definition of "reliable AI". European AI reliability tools could become the default compliance mechanism to standards requiring reliability guarantees. Once established, these standards favour the actors that shaped them.

3. Strategic Benefits from Leading on AI Reliability

Beyond the economic case, European leadership in AI reliability would deliver strategic benefits in terms of reduced dependence in strategic industries, talent retention, and leadership over a key part of the value chain. These considerations reinforce the economic arguments and provide additional justification for public investment.

a. Reducing dependence on foreign AI technologies in strategic industries

Safety-critical industries are also strategic for sovereignty. If solutions providing AI reliability guarantees and allowing deployment in these industries are not developed in Europe but elsewhere, Europe's technological dependence on foreign providers would increase further. European leadership in AI reliability would ensure these sectors are not dependent on foreign technology for safe AI deployment.

Even if domestic AI reliability technology does not become the leading global provider, its existence would reduce reliance from foreign technology when it comes to deploying AI in sovereignty-related sectors, since there would be a homegrown alternative available. In the more ambitious version of the research output - a safe-by-design homegrown frontier AI model - dependence would decrease even further.

Importantly, AI reliability guarantees would also increase model auditability, including of foreign AI models - rather than having to rely on opaque foreign technology, even if hosted locally. Ex ante reliability guarantees provide the auditability needed for technological control.

b. A talent repatriation and attraction lever

Europe faces a significant brain drain of AI talent. A publicly or philanthropy-funded non-profit (as currently suggested in the case of the *Frontier AI Initiative*) is unlikely to be able to compete with current industry-level pay packages. Targeted policies focused on visas and tax could help (and should be pursued). Here we argue that another important repatriation argument could be mission alignment[6].

A European R&I institution dedicated to improving AI reliability and offering (1) mission-driven work on guaranteed-safe European frontier AI, (2) quality of life advantages, and (3) research freedom, would be attractive to researchers aligned with the European approach to technology governance, even if it does not pay maximum compensation. For researchers who specifically want to work on and excel at researching

[6] There are precedents suggesting that researchers might prioritise impact over maximum compensation. For instance, despite not paying the highest compensation, Anthropic achieves 88% offer acceptance and 80% retention and poaches elite researchers from OpenAI and Google DeepMind – attributed partly to its founding safety mission. Further, the UK AI Security Institute successfully poached researchers from top labs with senior civil servant-level salaries. This suggests that a European institution with an explicit AI reliability mandate could attract researchers without paying maximum compensation.

ex'ante AI reliability, few opportunities exist at frontier companies. This is a specialised pool with strong intrinsic motivation. A European institution seriously tackling AI reliability would be uniquely attractive to these researchers. As argued in Section 7, high-risk high reward research cannot be done in academia alone.

In addition, the talent necessary to work on AI reliability is a partly-untapped talent pool. Indeed, AI reliability requires interdisciplinary expertise beyond machine learning, including e.g. formal methods or probabilistic programming. A European institute investing in AI reliability research could attract these other experts, most of whom are *not* currently working in AI companies.

Once established, a European center recognised as serious in reliability research would attract high-quality reliability-focused talent, who would then attract more high-quality talent, in a virtuous cycle. CERN and IMEC demonstrate these self-reinforcing dynamics, retaining excellent talent due to their prestigious reputations. Similarly, the deep learning revolution concentrated talent in a handful of early movers, such as OpenAI (2015), Anthropic (2021) and Google DeepMind (2014).

c. Leadership over an important layer in the AI value chain

Sovereignty does not require all parts of the stack to be local. Rather, sovereignty derives from managed interdependence through strategic leverage.

A leadership position in the AI value chain provides strategic leverage as:

- At a minimum, models developed elsewhere might rely on European expertise for deployment in critical sectors[7]. Even if model training takes place elsewhere, a European layer could become central to deployment. This would increase Europe's leverage in the global AI market.
- With ambitious enough investments (as discussed in Section 1), the research could lead to safe-by-design models developed in Europe, which might be favoured by deployers if European leadership in AI reliability is well established.

This lever is likely to be particularly powerful for deployment in sovereignty-related industries such as defence or critical infrastructure; this could increase Europe's bargaining position, as the technology it leads on becomes critical for other powers.

[7] See Section 2b for a discussion of why that might still be the case even if others catch up with the European innovation on AI reliability.

4. Europe's window of opportunity (***if*** it acts fast)

The economic and strategic case for European investment in AI reliability is strong. But the case is also urgent. This section argues that Europe can realistically secure leadership in AI reliability, provided it acts quickly, because the investment required is comparatively modest and the field remains neglected by frontier AI providers.

a. Leadership requires less investment than competing on scaling

AI reliability research costs a fraction of frontier model training. Hyperscalers ***each*** spend **$100-200 billion *annually*** on compute power infrastructure.

By contrast, **existing budget estimates** for an institute investing in high-risk high reward research on AI reliability **vary**, **but remain way below these annual costs, by several orders of magnitudes**.

The most comprehensive and most expensive *CERN for AI* proposal from the Center for Future Generations has a price tag of 35 billion over three years (including 10 billion for compute power). On a smaller scale, SaferAI recently produced an estimate for a *Frontier AI Initiative* focusing on AI reliability research, with a price tag of 64.5 million per year for the first three years of foundational research, including 8.5 million for the salaries of a team of 50 researchers[8] and 56 million on compute[9], and reaching 1.79 billion annually in the frontier model development phase, including 42.5 million on salaries[10], and 1.74 billion on compute[11]. Other estimates exist, some lower. The bottom line is that **AI reliability**

[8] Calculated on the basis of 12 senior research leads paid €120,000-300,000; 20 mid-career researchers paid €80,000-150,000; 18 junior researchers paid €55,000-100,000.

[9] According to AI reliability researchers in CAIA, the initial research is not compute-intensive but mostly involves theoretical development and many small-scale experiments to validate approaches. We estimate researchers would use a fifth of the compute available to frontier AI company researchers in this phase. OpenAI is estimated to have spent just over €4 billion on compute R&D in 2024. They had 770 employees in Nov 2023; assuming they more than tripled their headcount over 2024, and that at least ⅓ of their employees were researchers in 2023, they likely grew to about 750 researchers by the end of 2024. Assuming 7% (50/750) as many researchers as OpenAI in 2024, and a fifth the amount of compute per researcher, we reach €56 million.

[10] This is assuming a scaling to at least 300 researchers, including additional engineering talent, systems engineering talent for infrastructure, and expansion of core research team, compared to the foundational research phase. We estimate that this represents at least a 5x in cost, bringing the talent cost to 42.5 million annually.

[11] This is again calculated on the basis of OpenAI 2024 compute R&D expenditure, this time adding the cost of training GPT-4 - an extra 350 million - since the FAII would also be training at this point. Assuming half ⅖ of the staff (300/750), and the same amount of compute per researcher, this brings the total compute cost to ⅖ of their cost, e.g. 1.74 billion.

research is clearly a comparatively low-cost bet, relative to catching up on scaling approaches currently pursued by hyperscalers.

There are two reasons for this. First, **AI reliability research will be foundational at first**; **the main cost will be experimental compute and talent**, rather than the large-scale compute purchases required for training frontier models. To be clear, experimental compute is not trivial, companies do spend substantially on research compute (see Epoch AI, 2025). However, the scale is orders of magnitude smaller than training-scale investment. In the first phase, progress is more contingent on coordinating leading researchers under a single mission, which could be achieved if the *Frontier AI Initiative* became an agile and well-designed research institution, than on sheer compute scale.

Second, AI reliability is currently neglected by AI frontier companies (see Section 5), meaning that **the bar to become the leading investor in AI reliability is relatively low**.

b. Frontier AI providers are neglecting AI reliability, which creates an opportunity to move first

Research in the AI reliability agenda remains undersupplied by frontier AI providers, for reasons analysed in Section 5. Instead, current research on *ex'ante* AI reliability guarantees comes largely from nonprofits and universities, in an as-of-yet relatively uncoordinated fashion.

In this context, Europe's relative position behind the frontier of compute accumulation could allow it to be more agile in placing its bets, as less sunk cost allows flexibility to pursue alternative or complementary solutions without path dependency.[12]

5. Market Failure: Why Private Investment Will Not Close the Gap

The preceding sections have argued that AI reliability guarantees are a *sine qua non* condition for AI adoption in safety-critical sectors. This suggests the existence of a demand, from the safety-critical industrial leaders quoted above, and begs the question: if this demand exists, why are frontier AI companies not supplying it? This section identifies five structural reasons why the market is failing to deliver AI reliability research, and why public investment is therefore necessary.

[12] Importantly, this is not to suggest that Europe should not build more datacenters. Rather, it is to point out that lagging some distance behind the frontier reduces path dependency fixated on scaling, thus enabling flexibility in Europe's approach to developing frontier AI.

a. Competitive dynamics crowd out fundamental reliability research

In the current competitive environment focused on scaling AI models, frontier AI providers face intense pressure to release quickly. For instance, OpenAI has released a new model around every 3 months, and approximately every month a frontier AI model comes out, expanding the capabilities frontier.

Providers perceived as falling behind and risk losing investor confidence and talent to faster-moving competitors. While revenue increasingly matters, in the current phase of frontier AI development, investor sentiment remains heavily shaped by capability benchmarks and competitive positioning. Being 6 months behind on frontier AI capabilities can materially affect a provider's ability to attract capital, leading to intense competition.

By contrast, being 6 months behind on reliability and safety has no immediate commercial penalty. And the timeline of AI reliability research, which could take ~2 years to deliver results, does not align well with the intense pressure for short-term results.

The near-term focus of the current AI industry does not merely reduce the *quantity* of reliability and safety investment, it shapes the *type*. Frontier AI providers do employ substantial safety teams. However, the competitive environment pushes these teams toward approaches compatible with short-term capability gains and fast deployment cycles: RLHF, red-teaming, constitutional AI, and post-hoc evaluations. These are all *ex-post* safeguards, which often fail to be robust.

What competitive dynamics structurally disincentivise is the multi-year, foundational research needed for *ex-ante* reliability guarantees, research with uncertain timelines, which may require architectural changes, and which does not produce deployable improvements on a quarterly release cadence. One might ask: if frontier companies have hundreds of safety researchers, is the problem really under-investment, or is safety simply hard? We argue it is both, and that they reinforce each other. Safety research is indeed hard. But the competitive environment ensures that even the safety research that *does* get funded is oriented toward fast, incremental, deployable improvements, not toward the fundamental advances needed for formal certification.

A European institution pursuing *exante* reliability would not aim to do the same work better, it would do different work that competitive dynamics structurally prevent frontier companies from undertaking. We argue that this is a classic market failure in basic research: private firms underinvest in fundamental research even when they have large R&D budgets, because incentives push toward applied work with faster returns.

b. Frontier AI providers’ decisions are more investor-driven than demand-driven

Frontier AI providers currently depend more on attracting investors than on meeting demand for a particular product. Moreover, research in **AI reliability is less attractive to private investors than capability scaling**.

There are several reasons for this. First, **the returns of AI reliability research are not as predictable as that of scaling**. Scaling laws (Hoffmann et al., 2022) demonstrate a dependable relationship between compute and talent investment and capability improvement. Investors can thus expect returns with reasonable confidence. Conversely, AI reliability research is a high-risk, high-reward bet – the relationship between investment and outcomes is uncertain, compared to scaling. This means that with a short-term profit maximization logic, investors would prefer to invest in an established provider to continue scaling, rather than in a newcomer undertaking fundamental research. In other words, **AI reliability research is a less attractive bet than scaling when maximising private returns**.

Second, **AI reliability research has uncertain timelines to commercialisation**. AI reliability research is arguably currently at Technology Readiness Level (TRL) 1-3, requiring foundational advances in world models, safety specifications and verifiers (Dalrymple et al., 2024). As such, commercialisation timelines are likely ~2 years. This timeline might disincentivise private investment looking for quicker returns.

Therefore, AI reliability research is unlikely to be funded by private capital, despite the existence of a demand, notably from safety-critical industries.

For these reasons, we argue that supporting AI reliability research is an appropriate use of public R&I funding, allowing this area of innovation to overcome market failure until private investment can capture enough benefit to carry forward European champions. Without public intervention, private markets are likely to continue to undersupply foundational AI reliability research even as its value is clear.

c. Developers are focusing on error-tolerant applications, such as knowledge work

Frontier AI models perform well on tasks where occasional errors are acceptable. This includes drafting and search. These applications tolerate unreliable outputs because (a) human review is cheap, (b) errors are caught before deployment, and (c) the cost of individual mistakes is relatively low. Deployment of knowledge work-specialised LLMs is fast – allowing quick returns to be made, and unlocking a major business.[13]

Error-intolerant applications – such as aerospace, medical devices or energy systems – often require formal certifications, making AI deployment in these sectors less tractable and slower. Moreover, current models probably could not pass these certifications, explaining why developers prioritise deployment in error-tolerant rather than error-intolerant sectors.

d. Frontier AI providers face a sunk cost fallacy leading them to double down on scaling

Sunk costs in current talent and infrastructure are large. Providers have so far invested hundreds of billions in talent and infrastructure that is optimised for the scaling model, and for *ex-post* safeguards. Research in verifiably reliable AI could end up requiring fundamental differences to how models are trained, or move beyond the transformer paradigm altogether (for instance, neural network verification remains challenging with current methods, see e.g. Seshia et al., 2022).

While compute and individual researchers could in principle be repurposed, organisational shifts are not costless. The real switching costs lie in technical debt embedded in codebases and training pipelines optimised for current architectures; in organisational culture, workflows, and institutional knowledge oriented around the scaling paradigm; and in commercial relationships and product roadmaps that assume incremental capability improvements from existing approaches. These are not trivial to unwind, even for well-resourced organisations.

In addition, large investments sometimes lead to a lock-in effect due to a well-known behavioural economics phenomenon called the *sunk cost fallacy,* whereby actors can be reluctant to abandon a strategy because they have heavily invested in it, even when it would be more beneficial or rational to do so. This mechanism might be at play here, and explain AI providers' reluctance to research solutions which may eventually point to pivoting to different architectures (e.g. causal world models, neuro-symbolic systems), and hence to departing from their original bet on scaling.

e. Talent pipelines favour capability research rather than reliability research

Academic publications currently favour talent specialising in capabilities, rather than reliability. The dominant benchmarks and leaderboards in ML research measure task

[13] For instance, 13% of USA GDP was professional and business services in 2025, representing over USD$4 trillion.

performance (e.g. accuracy on ImageNet, scores on MMLU). Papers demonstrating state-of-the-art results on such benchmarks receive substantial attention from prestigious venues, whilst AI reliability has less obvious publication incentives. As a result, robustness in deep learning represented just over 1% of all papers at NeurIPS 2025.

In addition, AI reliability research requires interdisciplinary expertise. The skills required for AI reliability research span formal methods, probabilistic programming, and machine learning. Progress may therefore depend on coordination, grouping these researchers together to work towards a clear AI reliability mission - which is currently largely missing.

6. Existing European R&I Instruments Are Insufficient

The market failure identified above establishes the case for public investment. But can existing European R&I instruments fill this gap? This section argues that they cannot, and that the Frontier AI Initiative offers structural advantages that dispersed grants cannot replicate.

a. Horizon Europe is likely insufficient in scale and speed

Horizon Europe is likely the major viable existing instrument, but its AI funding may be too small and fragmented for the challenge. Horizon Europe allocated ~€2.6 billion to AI for 2021-2022; GenAI4EU plans ~€700 million across multiple programmes (European Commission, 2024). This funding is spread across many priorities (e.g. healthcare AI, robotics, virtual worlds) and hundreds of projects, rather than concentrated on AI reliability.

The scale mismatch with competitors appears stark. US hyperscalers are projected to invest ~$600 billion in AI infrastructure in 2026 alone (CNBC, 2026). A dedicated €500M-1B for AI reliability would likely represent an outsized investment in this specific domain; equivalent funding through Horizon Europe would probably be diluted across competing priorities.

Horizon Europe tends to fragment rather than concentrate effort. As explained above, the ARPA model's advantage lies in concentrating resources and talent on specific challenges under unified leadership (Azoulay et al., 2019). Horizon Europe's structure generally encourages many dispersed projects instead.

Horizon Europe's programming cycles may be too slow for AI. The 2026-2027 Work Programme was adopted in December 2025, with projects launching months later (IP Helpdesk, 2026). ARPA-style agencies can often deploy funding rapidly and pivot as circumstances change (Azoulay et al., 2019). For a fast-moving field like AI, this agility could be essential.

This is particularly concerning given that the window for leadership may be closing soon, requiring urgent action for bolstering Europe's competitiveness and sovereignty (see section 4).

b. The Frontier AI Initiative could offer structural advantages

Talent agglomeration under a unified mission

The Frontier AI Initiative could enable agglomeration of talent that distributed grant-based funding typically cannot achieve. Research breakthroughs often stem from cross-pollination between adjacent research streams. OpenAI's founding team in 2015 included researchers with expertise spanning deep learning, reinforcement learning and robotics, each leading research streams under a shared mission (Contrary Research, 2025). The Frontier AI Initiative could become a talent magnet for researchers interested in different strands of reliability research, offering a Bell Labs-like environment where bold ideas and deep collaboration are encouraged.

Academic research tends to be less interdisciplinary and more risk-averse due to institutional incentives. Research on scientific career incentives suggests that publication pressure discourages risk-taking: reward schemes that motivate effort inherently penalize researchers who generate unpublishable results (Gross & Bergstrom, 2024). As a result, scientists often choose projects that are safer than funders would prefer. Survey evidence indicates that academics believe journal ranking systems prevent them from conducting innovative and risky research, promoting instead work that is "safe, conforming, and mainstream" (Johann et al., 2024). **Hence, even funded research projects that are encouraged to take high-risk bets face a countervailing force which pushes towards more risk-averse approaches.**

Physical co-location may accelerate research in ways that distributed grants cannot. Bell Labs' success depended partly on physical proximity that facilitated spontaneous interactions and rapid iteration (Gertner, 2012). While Horizon Europe grants can in theory fund collaborative research, the distributed nature of academic work across institutions may limit the serendipitous exchange that characterises breakthrough-generating environments.

Long-term stability and adequate resources

Academic research often lacks adequate compute access. A recent survey of AI researchers worldwide found significant disparities between academic and industry scientists' access to computing power needed to train machine-learning models (Khandelwal et al., 2024). Research on the compute divide suggests that large-scale AI experimentation requires expertise in parallel and distributed computing that many academic groups lack (Besiroglu et al., 2024).

Project-based allocations may create too much uncertainty for long-term research. While allocating AI Factories compute to research groups could help, such allocations are typically project-based and uncertain by design. As explained above, the ARPA model tends to provide program managers with long-term budgets and flexibility to accomplish proposed programs (NCBI, 2003). This stability could enable 3-5 year research roadmaps that low-maturity research (e.g. safe-by-design AI) likely requires. DARPA also provides no-year money (i.e. funds that can be reallocated across years) and accepts unsolicited proposals, allowing researchers to pivot as opportunities emerge (Azoulay et al., 2019).

Agile, bureaucracy-light structures

The Frontier AI Initiative has the opportunity to create an agile, bureaucracy-free environment that enables curiosity-driven research. Existing R&I instruments often come with extensive reporting requirements, bureaucratic application procedures, and rigid conditions that limit the possibility to pivot research roadmaps when initial approaches fail.

Evidence suggests Horizon Europe's bureaucratic burden in particular remains substantial. The 2025 Interim Evaluation of Horizon Europe noted that European Partnerships face "complexity and bureaucracy" as open challenges (ERA-LEARN, 2025). Partnership managers have reported that bureaucracy is "hampering the partnership's biggest strengths, of speed and flexibility" (Science|Business, 2022).

Tolerance for failure and high-risk research

Academic career structures discourage high-risk research. Researchers who pursue ambitious projects that fail to produce publishable results damage their careers. Because unpublishable results could indicate lack of effort, incentive structures that reward publication inherently discourage scientific risk-taking (Gross & Bergstrom, 2024). Survey evidence suggests academics believe this pressure promotes safe, conventional research rather than ambitious work (Johann et al., 2024).

The Frontier AI Initiative could employ researchers on different terms. Program managers and researchers in ARPA-like institutions are evaluated on technical progress toward defined goals, rather than publication output. This removes a structural barrier to pursuing high-risk approaches. DARPA explicitly focuses on high-risk, high-payoff projects that other funders would likely reject. Horizon Europe grants, by contrast, are typically awarded to academics whose careers still depend on publications, preserving the underlying incentive toward safer research.

7. Institutional design for "high-risk, high-reliability" research

Having established both the need for public investment and the limitations of existing instruments, this section identifies the institutional features most likely to maximise the probability of research success.

The success of "ARPA-style" agencies appears to stem from a combination of institutional features that differ from traditional R&I funding Evidence from DARPA (Defense Advanced Research Project Agency) and ARPA-E (Advanced Research Projects Agency - Energy) suggests these features work as a system; adopting only some may fail to replicate their effectiveness (Azoulay et al., 2019). We identify four core features that a European ARPA for AI reliability could adopt.

a. Empowered program managers with fixed terms

Program managers are the central institutional innovation of the ARPA model. Unlike traditional grant administrators, program managers function as "risk-taking, idea-driven entrepreneurs heading up their own practice" (Azoulay et al., 2019). They conceive programs, select performers, and actively manage projects toward defined goals.

Critically, program managers serve fixed terms of three to five years (DARPA, 2025). This constraint creates urgency that permanent appointments may lack. Program managers must deliver results within their tenure, incentivising ambitious goal-setting and rapid iteration. The rotation also continuously refreshes the agency with new perspectives and expertise.

b. Active project management with milestone-based assessment

ARPA-style agencies structure programs in phases with explicit technical milestones. ARPA-E tracks progress against project milestones and may end projects that fail to meet goals so that remaining funds can be redeployed (ARPA-E, 2025). This differs from traditional grants, which often lock in research directions for the award duration.

The willingness to cancel underperforming projects distinguishes ARPA agencies from most public funders. It is a direct assurance against the sunk cost fallacy. Program directors frequently revise project milestones, budgets, and timelines based on technical progress (Azoulay et al., 2019). This active management resembles venture capital more than traditional grant-making, allowing resources to flow toward promising approaches and away from dead ends.

c. Flat organisational structure and minimal bureaucracy

DARPA maintains a flat organisational structure with only one management level between program managers and the agency director (National Academies, 2005). This enables rapid decision-making. Program managers can issue contracts and deploy funding 'practically overnight' using flexible contracting mechanisms (Azoulay et al., 2019).

This contrasts with the extensive reporting requirements and rigid conditions that often characterise traditional R&I instruments. A European AI reliability institute could be designed with similarly lean structures. Placing the institute outside existing bureaucratic frameworks (e.g. as an independent entity rather than a programme within Horizon Europe) could preserve the agility required for high-risk research.

d. Interdisciplinary collaboration under a unified mission

DARPA programs typically bring together academia, national laboratories, and industry under a single research mission (NCBI, 2003). This multidisciplinary teaming allows portfolios to combine basic and applied research with development and demonstration. The pairing of fundamental researchers with those focused on deployable products seems to accelerate technology transition.

For AI reliability, this approach seems particularly important. As noted in Section 5e), AI reliability research requires interdisciplinary expertise spanning formal methods, probabilistic programming, and machine learning. A European institute could convene these dispersed communities under a unified mission, creating the cross-pollination that breakthrough research often requires. Industry partnerships with European champions in safety-critical sectors (e.g. Airbus, Siemens, Alstom) would contribute domain expertise and ensure research priorities align with actual deployment requirements.

8. Complementary investment: Building a Reliable, Safe, and Secure AI Ecosystem

Alongside moonshots in verifiable AI reliability discussed above, we argue that Europe should invest in complementary research in safety and security innovations. This would help create a "reliable, safe and secure" AI ecosystem, thereby supporting a clear European branding and generating positive reinforcing dynamics between elements (e.g. security innovations would also support applications in safety-critical domains).

These complementary investments[14] in safety and security moonshots could initially concentrate on hardware security and verification mechanisms, secure data centers, and applications for cyberdefence.

a. Hardware security and verification mechanisms

Hardware-enabled security mechanisms embed verification capabilities directly into AI chips, allowing regulatory enforcement and cross-stakeholder accountability at the physical layer. This includes secure enclaves for model weights, cryptographic attestation of compute usage, and hardware-rooted enforcement of safety evaluations - see the FlexHeg agenda, as example.

This could be done by leveraging Europe's existing strengths in hardware security research, such as COSIC at KU Leuven, TU Graz, or start-ups like Amodo design, and a strategic relationship with ASML. Hardware verification research does not necessarily require fabrication capabilities, it requires specification expertise that can later be implemented by manufacturing partners. European researchers can define what security properties chips should enforce, even if fabrication occurs elsewhere.

Investing in hardware development at scale would require semiconductor manufacturing capabilities that Europe currently lacks. In addition, hardware-level guarantees must be paired with model reliability advances to provide assurances able to unlock the full economic and strategic leverage gains discussed above. This makes it a really good complementary investment, alongside efforts to improve model reliability. Hardware security can provide a root of trust at the hardware level complementing model-level reliability innovations in establishing a European "verifiably reliable AI" ecosystem.

b. Secure datacenters

Datacenters handling AI models face sophisticated threats including model theft, weight exfiltration, and supply chain attacks. The RAND Corporation's datacenter security framework identifies five security levels, with current frontier labs operating at approximately Level 2-3. Reaching the maximum Level 5 (i.e. resilience against state-level adversaries) requires substantial investment.

Datacenter security builds on established disciplines (physical security, access control, cryptographic protocols) where Europe has mature expertise. The challenge is primarily capital investment and actual implementation by datacenters' operators rather than unsolved research problems. European cloud providers (e.g. OVHcloud or Deutsche Telekom) and the planned AI Gigafactories provide a foundation for implementing higher

[14] Note that these complementary investments should come *on top of* other efforts in domestic compute build-out and talent repatriation policies, which remain necessary.

security standards to intensify and expand secure AI implementations in the EU. Further, the EU's existing regulatory tools for critical infrastructure protection (e.g. the NIS2 Directive) provides a framework that can be extended to AI datacenters.

Guaranteeing the highest infrastructure security level would be complementary to core investments in verifiably reliable AI models. It would provide a solution to unattended challenges, such as weight model security against attempts of theft or manipulation. It would also contribute to the overall "reliable, safe, secure" European AI branding.

The objective of developing secure datacenters could be integrated into the AI gigafactories initiative. This would benefit the entire ecosystem and support other EU programmes, such as the *Frontier AI Initiative*.

c. AI for cyber-defence

Alongside moonshots in verifiable reliability, Europe could invest in AI applications that directly enhance safety and security outcomes, with cyber-defence as a particularly high-leverage target. This would strengthen Europe's ability to deploy AI safely in critical sectors, where cyber risk is already a binding constraint.

While reliability research reduces risks from model behaviour; cyber-defence reduces risks caused by the systems surrounding the model (risk such as e.g. model theft, supply-chain attacks, data poisoning). Again, investing in defensive AI therefore reinforces the whole "Reliable, Safe and Secure AI" package.

Europe could for instance prioritise research into Security Operation Centre (SOC) augmentation for critical infrastructure operators (e.g. triage, alert summarisation, incident reporting, and guided containment with strict human-in-the-loop controls); phishing and social-engineering defence at scale (detection of targeted campaigns and defensive training/simulation tooling for high-risk workforces); or vulnerability and patch prioritisation for OT/ICS environments (helping operators focus scarce resources on the most exploitable and safety-relevant risks).

Appendix A: estimating the size of "safety critical" industries in the European economy, compared to the US.

European estimate

To create a measurable proxy for the "safety-critical sector", we define a basket of industries from the NACE Industry Classification Codes for which Eurostat data is available. This basket includes: chemicals manufacturing, pharmaceuticals manufacturing, aerospace, defence, electricity & gas / utilities, rails & pipelines, air transport, and warehousing and support activities for transport. Using NACE Rev. 2, these industries are proxied using the following variables:

Industry	Eurostat proxy variables
Petroleum and gas refinement	C19, "Manufacture of coke and refined petroleum products" *This includes the transformation of crude petroleum and coal into usable products (petroleum refining through such techniques as cracking and distillation) as well as the manufacture of gases such as ethane, propane and butane as products of petroleum refineries.*
Chemicals manufacturing	C20, "Manufacture of chemicals and chemical products"
Pharmaceuticals manufacturing	C21, "Manufacturing of basic pharmaceutical products and pharmaceutical preparations"
Aerospace, defence & other transport manufacturing	C30, "Manufacture of other transport equipment". *This includes building of ships and boats, manufacture of railway locomotives and rolling stock, of air and spacecraft and related machinery, of military fighting vehicles.*
Utilities	D, "Electricity, gas, steam and air conditioning supply" + E "Water supply; sewerage, waste management and remediation activities"
Land transport (passenger and freight)	H49, "Land transport and transport via pipelines". *Note that this includes the transport of passengers and freight via road and rail, as well as freight transport via pipelines.*
Air transport	H51, "Air transport"
Transport logistics	H52, "Warehousing & support activities for transportation". *This corresponds to e.g. operating of transport infrastructure (e.g. airports, harbours, tunnels, bridges, etc.), the activities of transport agencies and cargo handling.*

Note that this basket is designed as a structural proxy for the relative concentration of safety-critical industries across economies, not as a precise estimate of the market for AI reliability guarantees. Not all economic activity within these sectors requires formally certified AI: for instance, transport warehousing logistics likely includes applications where current AI could be deployed without formal certification. The relevant question is not whether 100% of GVA in these sectors would be unlocked by reliability guarantees, but whether Europe's greater structural concentration in these sectors means it would benefit proportionally more from advances in AI reliability than economies (such as the US) that are relatively less concentrated in these industries. We argue this structural comparison holds even under conservative assumptions about which sub-sectors and applications would require formal certification.

In 6 countries (Ireland, Lithuania, Luxembourg, Malta, Poland, and Sweden), some of the data is classified as confidential. To avoid biases, we exclude these countries and produce an EU-21 measure, based on data in: Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Italy, Latvia, Netherlands, Portugal, Romania, Slovakia, Slovenia, and Spain.

The table below shows the total share of the gross value added across the basket of "safety-critical" sectors, across the 21 countries for which complete data is available, in current prices in the 21 countries' total gross value added.

Share of "safety critical industries" in the economy, EU-21, 2023 (€ billion)

Sector label (NACE Rev.2)	**GVA 2023 (€bn)**	**Share of the 21-country total GVA**
Petroleum & gas refinement (C19)	33.1	0.24%
Chemicals manufacturing (C20)	133.8	0.97%
Pharmaceuticals (C21)	124.3	0.90%
Aerospace and defence manufacturing (C30)	66.7	0.48%
Utilities (D+E)	427.6 (349.4 + 123.2)	3.43%
Land transport (passenger & freight) (H49)	270.2	1.96%
Air transport (H51)	36.4	0.26%
Transport logistics (H52)	264.7	1.92%
Total "safety-critical" industries	**1,401.8**	**10.17%**
21 countries total GVA (denominator)	13,779.4	100%

Source: Eurostat, dataset "nama_10_a64 — Gross value added and income by detailed industry (NACE Rev.2)" (indicator B1G, unit current price million EUR, annual - converted in billion above).

Notes: The data displayed is gross value added (GVA), not GDP at market prices, which is equivalent to GVA + taxes less subsidies on products.

This is an operational basket based on best available proxy in available national accounts data at NACE A*64 detail level: for instance, C30 ("other transport equipment") is broader than "aerospace & defence". H52 is broad transport logistics, not only safety-critical logistics.

2023 values are marked "provisional (p)". Revisions are possible as Eurostat updates the national accounts tables.

United States estimate

For the US estimate, we use data from the Bureau of Economic Analysis on annual GDP/industry (using NAICS codes), in annual, current-dollar. BEA "GDP by industry" statistics measure industry value added and are therefore directly comparable to GVA-based sector shares. Mapping US data onto the EU basket defined above, we use the following variables:

Industry	Eurostat proxy variables	BEA proxy variables
Petroleum and gas refinement	C19, "Manufacture of coke and refined petroleum products" *This includes the transformation of crude petroleum and coal into usable products (petroleum refining through such techniques as cracking and distillation) as well as the manufacture of gases such as ethane, propane and butane as products of petroleum refineries.*	NAICS 324, "Petroleum and Coal Products Manufacturing"
Chemicals manufacturing	C20, "Manufacture of chemicals and chemical products"	NAICS 325, "Chemical manufacturing" *Note that this includes pharmaceuticals*
Pharmaceuticals manufacturing	C21, "Manufacturing of basic pharmaceutical products and pharmaceutical preparations"	
Aerospace, defence & other transport manufacturing	C30, "Manufacture of other transport equipment". *This includes building of ships and boats, manufacture of railway locomotives and rolling stock, of air and spacecraft and related machinery, of military fighting vehicles.*	NAICS 3364, "Aerospace Product and Parts Manufacturing" + NAICS 3365, **"**Railroad rolling stock manufacturing" + 3366, "Ship and boat building" + 3369, "Other transportation equipment manufacturing"
Utilities	D, "Electricity, gas, steam and air conditioning supply" + E "Water supply; sewerage, waste management and remediation activities"	NAICS 22, "Utilities" + NAICS 562, "Waste Management and Remediation Services"

Land transport (passenger and freight)	H49, "Land transport and transport via pipelines". *Note that this includes the transport of passengers and freight via road and rail, as well as freight transport via pipelines.*	482, "Rail" + 484, "Truck" + 485, "Transit & ground passenger" + 486, "Pipeline"
Air transport	H51, "Air transport"	NAICS 481, "Air transportation"
Transport logistics	H52, "Warehousing & support activities for transportation". *This corresponds to e.g. operating of transport infrastructure (e.g. airports, harbours, tunnels, bridges, etc.), the activities of transport agencies and cargo handling.*	NAICS 493, "Warehousing & storage" + NAICS 488, "Support activities for transportation" ***(estimated)****

* Note: In the BEA "GDP by State" industry data which we had access to, NAICS 488 is not available as a standalone value-added series for the United States. Instead, it appears inside a bundled industry category: "Other Transportation and Support Activities (NAICS 487–488, 492)". This bundle includes: 487, "Scenic and sightseeing transportation", 488, "Support activities for transportation", 492 "Couriers and messengers". Only 488 corresponds to the EU H52 "support activities" concept. NAICS 487 and 492 do not correspond to H52 and therefore should not be fully included.

We estimated the share of 488 within the bundled category using the US Census Bureau quarterly revenue series. This data provides an observable, recent measure of the relative size of 488 within the bundle.
We computed the relative revenue shares of 487, 488, and 492 using recent multi-quarter totals, as follows: Share of 488 = Revenue488 / (Revenue487 + Revenue488 + Revenue492). We found that 67.6% of the bundle is attributable to NAICS 488.

We then applied that proportion to the 2023 GDP (value added) of the bundled industry, e.g. VA 488 = 0.676* VA (488+ 487 + 492). This produced an estimated 2023 value added for NAICS 488.

The US proxy for the H52 Eurostat variable is therefore the sum of the observed NAICS 493 and the reconstructed estimation of NAICS 488.

Share of "safety-critical industries" in the economy, US, 2023 (billion US$)

Sector label (NAICS)	**Value added 2023**	**Share of US total value added**
Petroleum & gas refinement (NAICS 324)	218.0	0.78%
Chemical & pharmaceuticals manufacturing (NAICS 325)	514.6	1.85%
Aerospace, defence & other transport manufacturing (NAICS 3364–3366, 3369)	190.8	0.69%
Utilities (NAICS 22 + NAICS 562)	540.3 (457.8 + 82.5)	1.95%
Land transport (passenger and freight) (NAICS 482 + 484-486)	427.1	1.54
Air transportation (NAICS 481)	168.3	0.61
Transport logistics (NAICS 493 + estimated 488)	260.1	0.94%
Total "safety-critical" industries	**2,319.2**	**8.34%**

Total US total value added (denominator)	27,812	100%

Source: U.S. Bureau of Economic Analysis, dataset “Gross Domestic Product by State, annual, current dollars, United States”. Denominator (All industry total, 2023) = 27,811,517 (million US$).

Comparability caveats

- NAICS and NACE do not align perfectly. This method reduces distortion but does not eliminate structural differences.

Appendix B: estimating the share of “industry champions” operating in safety-critical industries in Europe, China and the United States

In order to proxy the share of respective EU, US and China “champions” in safety-critical industries, we estimate the share of each region’s large/mid-cap publicly listed corporate firms operating in safety-critical sectors. The output is a percentage of equity market capitalization in those sectors, within a large/mid-cap regional index (where “champions” can usually be found).

To do so, we look at MSCI indexes, since these are broad large/mid cap regional indexes. These indexes are market-cap weighted: the biggest firms drive the weights, which make them adequate to estimate the size of “champions”.

Specifically, we look at:

- Europe: MSCI Europe Index - Index Factsheet (Jan 30, 2026)
- US: MSCI USA Index - Index Factsheet (Jan 30, 2026)
- China: MSCI China Index - Index Factsheet (Jan 30, 2026)

Each of those factsheets includes a section on “sector weights”, with % weights by sector. To proxy “safety-critical sectors”, we the following sector listed in the factsheet:

- “Materials”, which includes e.g. chemical manufacturing
- “Industrials”, which includes aerospace & defense and transportation services
- “Health Care”, which includes companies involved in pharmaceuticals & biotech
- “Energy”
- and “Utilities”

See the MSCI Global Industry Classification Standard (GICS) for sector definition.

From the “sectors’ weights” section in the factsheet sourced above we pull the following data:

Industry	MSCI Europe	MSCI USA	MSCI China
Materials	5.32%	1.98%	5.65%
Industrials	19.39%	8.85%	4.75%

Health Care	13.96%	9.51%	4.68%
Energy	4.23%	3.21%	2.81%
Utilities	4.78%	2.21%	1.74%
Total	47.68%	25.76%	19.63%

Important caveats:

- This is an equity-market-cap share, not an “economy share.” Sector weights reflect valuations, not production, revenue, employment, or value added.
- These estimates cover only publicly listed companies in the index universe. Private firms, state-owned entities not listed, and smaller firms outside the index are excluded.
- Europe in MSCI Europe is not the same as EU-27. MSCI Europe includes developed European markets (including the UK and Switzerland).
- Sector-level mapping is coarse. “Industrials” includes many things beyond aerospace/defense and regulated transport; “Health Care” includes more than pharma; “Materials” includes more than chemicals.

SaferAI

About SaferAI

At **SaferAI**, our AI risk management expertise uniquely positions us at the intersection of technical research and policy. We are a founding member of the US AI Safety Institute Consortium, a founding member of the Hiroshima AI Process Friends Group Partners' Community, and we regularly contribute to OECD/GPAI Expert Community meetings. We developed the first public rating system for AI companies' risk management practices and published a well-established risk management framework for advanced AI.

🌐 safer-ai.org ✉ contact@safer-ai.org 👍 LinkedIn | X