What we talk about when we talk about AI for Science

Arq Foundation

Bengüsu Özcan

Senior Researcher, Frontier AI Governance

Contents

Executive summary

AI for science is a strategic investment priority across the European Union, the United Kingdom, and the United States. These large funding commitments are essential to sustain scientific progress. But realising the value of AI for science investments will depend on two key factors that existing strategies sometimes miss. First, "AI for science" is not just one thing. The term covers at least two distinct layers of the scientific research ecosystem, which solve different problems and depend on different types of inputs and investment. Second, because frontier AI capabilities change so fast, where they create value in science will keep shifting; and these opportunities cannot be identified by an investment strategy that is set once and designed to run for years. Instead, investment strategy needs to be agile, offering capacity to try new tools, placing early bets, and re-targeting investment as the landscape shifts.

These two factors point in a clear direction. AI for science investment should identify the distinct layers where AI creates value and target each one specifically, guided by a dynamic roadmap that concentrates resources on the highest-value capabilities and the bottlenecks blocking them. If investments are not targeted this way, we risk spreading resources too thin instead of building lasting, flexible scientific capacity – especially in Europe, where research funding is closely tied to competitiveness and strategic autonomy.

Today, two high-impact opportunities occupy distinct but complementary layers of the scientific AI stack. They are often discussed together, but they solve different problems, depend on different inputs, and mature on different timelines.

The first layer, **scientific foundation models,** are designed to model natural phenomena directly. For example, AlphaFold can predict protein structures at scale, and GraphCast can generate high-resolution weather forecasts in seconds. These models unlock capabilities that simply did not exist before, often at dramatically lower cost, and once built, they function like shared scientific infrastructure.

The second layer is **AI scientist systems.** Rather than modelling nature, they aim to organise and run parts of the scientific process. These systems are typically built on general-purpose AI models, with additional structure that allows multiple agents to work on scientific questions in parallel. They run structured workflows where literature review, hypothesis generation, and digital tools are used to refine the model's final output; and they are designed to keep human scientists in the loop. Some are offered by specialised developers, such as Kosmos by Edison Scientific, while others are built by frontier AI companies who offer general-purpose models as well, such as Google DeepMind's Co-Scientist and OpenAI's GPT-Rosalind. AI scientist systems' promise lies not in depth, but in breadth: they enable scientists to explore ideas faster and more systematically. If successful, they could compress research timelines from years to months, a shift large enough that only institutions with access to such systems would remain competitive in producing scientific knowledge.

These two layers of "AI for science" are complementary but require fundamentally different investment strategies. Scientific foundation models depend on large volumes of domain-specific data, typically generated by national observatories or long-running research programmes. The bottlenecks for these models tend to be coordination problems, such as building structured data pipelines. AI scientist systems face a different constraint: there is very little structured data on how science is actually done, including how scientists make judgements under uncertainty or learn from failed attempts. Because scientists publish final outputs, such as papers and datasets, rather than the reasoning and iteration behind them, this kind of workflow data that AI scientist models need to learn from is largely missing.

These two layers call for different investment approaches. The data needed to build or adapt each layer is fundamentally different. AI scientist systems in particular risk being overlooked: they are newer, and scientists do not typically collect and share the data that they need. Investment strategies also have to enable both development and procurement. Some leading AI systems are developed by private companies, so securing access to them matters as much as developing new capabilities at home. As the boundary between the layers shifts and frontier capabilities advance rapidly, AI for science investment strategy will need to be revisited as a living roadmap: it should provide funding across both layers, identify and target bottlenecks precisely, and build in a mechanism to collect ongoing evidence in order to adjust the strategy.

This is especially pressing for Europe right now: a large EU commitment through the Resource for AI Science initiative is already in place, and Framework Programme 10, Europe's flagship research and development investment, is still being shaped. We recommend that science and R&D investors, and EU policymakers in particular, should:

Distinguish the distinct layers in AI for science, and develop funding mechanisms suited to each. These layers depend on different data, expertise, and infrastructure, so the binding bottlenecks will differ between them and should be targeted directly. Some funding mechanisms will mean building capability at home, others securing access to leading systems developed elsewhere, and they should stay flexible enough to do both. Public funders could also explore alternatives for cases where conventional landscape assessment and funding application processes move too slowly, such as partnering with faster intermediaries like philanthropic organisations that scout opportunities in specific areas.

Treat the strategy as a living roadmap, not a one-off allocation. The AI frontier moves fast, so where it adds value will keep shifting. In practice, this means reviewing the portfolio regularly, letting promising bets scale, and sunsetting stalling initiatives as new evidence comes in.

Monitor and test the scientific capabilities of general-purpose AI. These capabilities are built both by specialised developers and, increasingly, by frontier AI companies. Scientific AI and general-purpose AI are not separate worlds, and strategy should track how advances in frontier models reshape what is possible in science.

Give scientists a structured way to test these systems. Most researchers have little experience judging where AI scientist systems work or fail. In practice, this means creating capacity for scientists to run experiments with new tools. These experiments should generate clear audit trails

and lessons learned should be shared, so that findings feed back into the broader AI for science strategy.

Recognise that access will be as decisive as capability. Some scientific AI systems are offered only by private companies or require in-house engineering to build, so access depends heavily on procurement conditions and available resources. Europe should build its own AI for science capabilities while ensuring researchers can access leading existing systems through smart procurement, interoperability standards, and partnership terms that avoid lock-in. Falling behind in access can be as damaging as falling behind in development.

AI is not a silver bullet for scientific discovery today. Different fields face unique bottlenecks that AI advancements alone cannot resolve, from foundational limits to regulatory hurdles. However, scientific AI capabilities pose one of the most consequential open questions about how the scientific process itself might change. Scientific foundation models are already a proven pillar of AI-enabled science. AI scientists, by contrast, are first-generation systems whose value can only be determined by deliberate experimentation and learning by building. The more precisely and dynamically Europe targets its AI for science investments across these layers, the more likely it is to remain a leading contributor to global scientific progress.

Introduction

AI for science has emerged as a policy priority across the world, reflecting a broader recognition that advances in AI could alter the pace, structure, and direction of scientific discovery itself. In Europe, the momentum was visible in 2025 with the launch of Resource for AI Science in Europe (RAISE) [1]. Announced by the European Commission, with an initial €107 million allocation from the Horizon Programme for 2026–27, RAISE aims to enable AI-driven scientific research, with strategic enablers such as providing researchers access to substantial compute as a fundamental need for training and deploying advanced models.

Looking beyond Europe helps clarify what is at stake. While RAISE emphasises coordination and capacity-building within an existing research framework, developments in the United States imply that the U.S. has a more explicit ambition to use AI to reshape the scientific process itself. This is most clearly expressed in the U.S. Genesis Mission [2], launched in November 2025. Framed as a new "Apollo Project" for science, Genesis positions AI as core scientific infrastructure, and focuses on integrating national laboratory data with shared, AI-enabled platforms across labs, universities, and industry.

These two initiatives are not directly comparable in terms of scale, mandate, or intended scope. However, as RAISE infrastructure continues under FP10 – the next multi-year European Union research and innovation Framework Programme following Horizon Europe – AI for science is likely to become a structural component of Europe's research strategy over the coming decade.

"How to do science better" is a well-established policy priority: few public investments rival science in their long-term economic [3] and social [4] returns. Knowledge generated in one field or country spills [5] across sectors, borders, and time, making science and research and develop-

ment (R&D) one of the few policy levers with consistently high payoffs [6]. So far, AI has been one of the enablers of this logic: AlphaFold's Nobel-recognised breakthrough in protein folding is the canonical example of AI accelerating [7] drug discovery, but AI also has a visible track record [8] of enabling more creative, cross-disciplinary science in other areas. Today, this pattern is not changing, but AI's reach could be expanding. If AI can compress research cycles and uncover patterns beyond human cognitive limits, problems that have resisted decades of effort – from room-temperature superconductors to drugs that extend healthy lifespans – could become far more tractable. In this case, AI would act as a powerful amplifier of one of the most socially valuable activities that governments already support: scientific research and discovery.

The challenge is that the rapidly evolving AI-in-science landscape is outpacing policymakers' shared understanding of the area. New models, infrastructures, and institutional arrangements are emerging, often faster than policy frameworks and scientific best practices can adapt. When setting policy and funding priorities, clarity about which AI systems meaningfully advance science is just as important as technical progress.

This report offers a structured way to think about where strategic investments in AI for science should be targeted. First, it explores two high-stakes types of AI systems in science – scientific foundation models and AI scientist systems – and shows how they differ in architecture, data requirements, cost, and their role in the research ecosystem. Second, it examines whether frontier general-purpose models are likely to absorb these capabilities, drawing on recent evidence. Third, it translates these distinctions into policy implications, with particular attention to where European investment and coordination can be most precisely targeted.

AI for science means different things

"AI for science" is often discussed, but it refers to a wide range of systems with very different roles, requirements, and implications. This has made the term imprecise, particularly for funders and policymakers who need to assess different AI for science pathways and decide which ones warrant long-term investment.

To address this ambiguity, a growing body of work has proposed taxonomies to map how AI is being used across scientific fields and practices. These frameworks do not claim to point to strict boundaries, but they help surface patterns, showing where AI is already delivering value and where it remains experimental.

These taxonomies are useful for capturing the breadth of activity grouped under "AI for science", but they do not translate into strategic judgement for policy and public investment. Such decisions require a big-picture view of how AI systems shape scientific capacity over time and create long-term dependencies, which is shaped by both public and private initiatives.

Among recent public initiatives, the EU's RAISE [17] and the U.S. Genesis Mission [18] are two prominent examples. Both are multi-layered programmes that fund and coordinate many different things at once — dataset creation, infrastructure, partnerships — so neither maps neatly onto a single approach. But in their current shape, both lean heavily toward creating and enabling the use

How the research literature maps “AI for science”

One class of taxonomies organises AI by scientific domain and task. Frameworks such as Nature AI for Science 2025 [9] and the Royal Society’s discipline-oriented mappings [10] classify systems by fields they support (e.g. chemistry, biology, materials…) and by use cases such as reaction prediction or climate downscaling. These maps are useful for identifying discipline-based best practices and needs, such as where AI adoption is fastest, or where bottlenecks persist due to infrastructure gaps or data scarcity.

A second class of frameworks maps AI onto the scientific process itself. The Royal Society’s Science in the Age of AI report [11] charts how AI interacts with core methodological pillars of the scientific process, including modelling and simulation, experiment design, or reproducibility. This framework treats AI less as a collection of domain-specific tools and more as an infrastructure layer reshaping scientific practice. It also discusses how systemic issues in science intersect with AI, for example reproducibility challenges, widening skill gaps, and shifting incentives where “being good at AI” risks overshadowing “being good at science”. Some studies [12] offer more integrated mappings between the two classes of taxonomies mentioned so far, such as classifying AI applications per discipline, but also identifying cross-cutting tasks that similar AI capabilities are used for across disciplines.

A third, more recent strand of research focuses on AI systems that attempt scientific reasoning and autonomous research. Surveys [13] of LLM-based agentic systems (e.g. Google’s Co-Scientist, Sakana AI's AI Scientist) and hybrid human–AI workflows examine their scientific capabilities from generating hypotheses to planning experiments. Other studies help chart this new territory by clustering [14] models based on how much human oversight is required, or by their design foundations [15]. These studies reveal a mixed picture: AI scientists can accelerate experimental design and propose research directions at scale, yet they can still show weaknesses, such as producing shallow hypotheses or performing unevenly across disciplines. As the field evolves rapidly, systematic studies lag behind, since newer models [16] may not fit neatly into existing classifications.

of large, domain-specific datasets that require immense amounts of institutional coordination, public-private partnerships, and large amounts of compute access. This is typically the type of data and infrastructure needed for scientific foundation models, which we describe in the following section.

In contrast, industry tends to emphasise a different scientific AI capability. Companies such as OpenAI [20], Anthropic [21], and Google DeepMind [22] which develop large general-purpose models all target scientific discovery as a core use case they want their models to get better at, while companies such as FutureHouse [23] focus solely on scientific discovery capabilities via building dedicated AI scientist systems. The goal here is not just to model the world but to support the scientific discovery process itself. This ambition shows up in U.S. public framing [19] too, but it is far more pronounced in the private sector than in existing public initiatives.

If we look at where sustained public and private commitments are already concentrating, we see that investments focus on two dominant, high-stakes categories of AI: scientific foundation models, which are designed to model natural phenomena, and AI scientist systems, which aim to automate and coordinate parts of the scientific research process. Automated laboratories also have clear potential to reshape scientific capacity and are also highlighted in the taxonomies we examined, but their physical nature introduces a distinct set of dependencies to analyse. For the purposes of this report, we leave them aside and focus on these two software-based layers.

These layers share close similarities and, as such, are the most prone to conceptual confusion. Both will require highly resource-intensive investments, and they are among the approaches most likely to reshape how scientific knowledge is produced and used across fields.

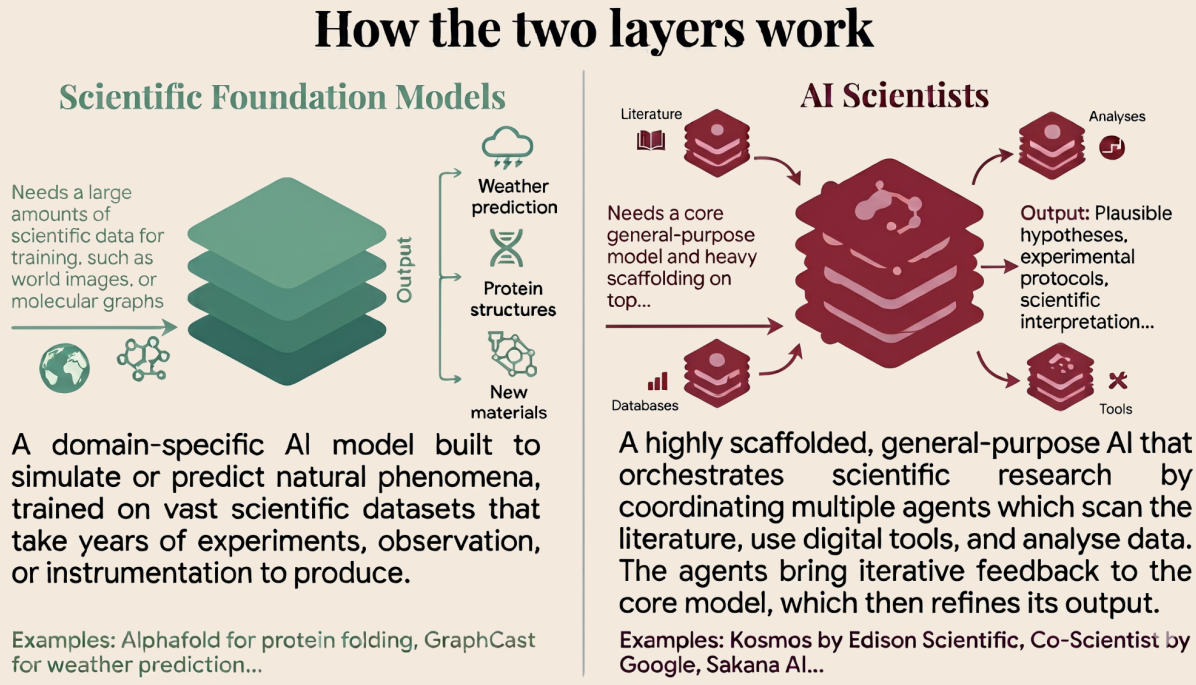

Figure1: Even though they are both AI models typically trained on large data sets, scientific foundation models and AI scientists have different underlying working mechanisms.

A strategic distinction: Modelling nature vs. automating discovery

Scientific foundation models are oriented towards representing the natural world itself, whereas AI scientist systems aim to organise and accelerate the process of discovery. Focusing on these two strands makes the strategic trade-offs clearer: both types of AI function as shared scientific capacity – capabilities that extend beyond individual projects and shape entire research ecosystems – but they do so in fundamentally different ways. Understanding how these approaches differ, and how they interact, is essential for assessing where investment, governance, and coordination matter most.

Modelling nature: Scientific foundation models

WHAT THEY ARE

Scientific foundation models [24] are designed to model natural phenomena and physical processes. Typically domain-specific, they learn the underlying dynamics of complex systems, enabling insights that would be computationally intractable or experimentally impossible through traditional methods. Rather than making research workflows faster and more efficient, they open new doors to natural phenomena at scales and resolutions inaccessible to human perception or existing instruments.

> FourCastNet [25] was initially released in 2021 and continuously developed by NVIDIA in collaboration with public research institutions such as Lawrence Berkeley National Laboratory. It is a weather forecasting model that learns atmospheric dynamics directly from decades of historical weather data. It generates short-term forecasts efficiently, supporting rapid decisions that need to be made under high uncertainty.
>
> GraphCast [26], developed by Google DeepMind in 2023, addresses the same core task of global weather prediction and produces accurate 10-day forecasts in under a minute, outperforming the European Centre for Medium-Range Weather Forecasts (ECMWF), the long-standing gold standard. It is part of a broader set of publicly available scientific foundation models from DeepMind, including AlphaFold in protein structure prediction and GNoME in materials discovery.Comparative studies [27] find that GraphCast performs better at longer forecast horizons, enabling earlier warnings for extreme events, while FourCastNet performs particularly well at rapidly generating weather scenarios (ensembles) for shorter time frames. These complementary AI systems, optimised for different operational needs, thus expand meteorological agencies' toolbox

THE DATA THEY NEED

These models are trained on specialised scientific data – molecular graphs, spatiotemporal fields, structured experimental measurements, or high-dimensional sensor outputs. These datasets differ fundamentally from the text or multimodal corpora used to train LLMs. Such data is often expensive to generate, as it can only be obtained by specific experimental setups or observational

instruments, such as the Large Hadron Collider at CERN or the James Webb Space Telescope. Converting large, messy bodies of raw observations into structured and labelled datasets requires substantial domain expertise and can be highly labour-intensive. While this data can sometimes be made public, availability often depends on sensitivity and ownership. Even when public, these datasets can be fragmented across institutions, so coordination is required to plug them into AI pipelines.

> AlphaFold [28] was trained on approximately 170,000 protein structures from the Protein Data Bank and billions of protein sequences via databases like UniProt. While this data is fully open to the public, it was accumulated through decades of scientific effort across the world.
>
> Morpheus [29], an AI model that automatically identifies and labels galaxies and their structures, was built using data from the James Webb Space Telescope, developed through a federal research programme that cost at least $1.6 million. In comparison, the James Webb Space Telescope [30] itself cost roughly $10 billion and took over 25 years to develop, generating [31] about 235 GB of space imagery that cannot be replicated without this unique infrastructure. This data becomes public after a 12-month proprietary period, and processing it requires substantial computational resources, effectively limiting the training of models like Morpheus to well-resourced institutions.

HOW THEY ARE TRAINED

The underlying architecture of scientific foundation models varies widely, reflecting the structure of their training data and the systems they study. For example, molecules are often represented [32] as graphs, so AI models trained on this data for the purpose of new material discovery often use graph neural networks, which are related to but distinct from the transformer architecture used in LLMs. Scientific foundation models[1] are trained and evaluated using methods grounded in established scientific theories, which check whether outputs obey physical laws and match experimental observations. While these models are not perfectly accurate, this grounding in verifiable principles offers [33] more validation than is available for most general-purpose model use cases, which tend to be inherently messier and more context-dependent. These scientific models also require collaboration between domain scientists, who define what "correct" means in their field, and machine learning researchers, who translate that knowledge into model architectures.

> AlphaFold uses its novel Evoformer architecture: 48 stacked processing layers that learn from patterns across thousands of related protein sequences. [2]
>
> GraphCast employs a graph neural network with 36.7 million parameters that reads in weather conditions, refines them through internal computations, and outputs predictions across a multi-layered representation of Earth's surface.[3]

1. While "foundation model" is often used to describe general-purpose systems such as large language models, the term is also used in the literature for models that learn broadly reusable representations within a given domain (e.g. astronomy, genomics, climate). This report adopts the latter usage when referring to scientific foundation models.
2. The description of the Evoformer architecture and its 48 stacked blocks comes from the "Methods and architecture" overview in Jumper et al. (2021), which details how multiple sequence alignments are iteratively processed to predict protein structure.
3. The encode–process–decode GNN design, parameter count (36.7M), and multi-mesh Earth representation are reported in Lam et al. (2023), which outlines the model architecture used for medium-range global weather forecasting.

WHAT IT TAKES TO ACCESS AND USE THEM

Scientific foundation models generally require significant compute resources to train, but far less than frontier general-purpose models, which dominate public discourse. Models such as AlphaFold and GraphCast were trained over days or weeks using large supercomputing facilities comparable to those under EuroHPC [34], representing a meaningful but contained infrastructure investment. Once trained, they are inexpensive to run: AlphaFold2, for example, typically takes 5–30 minutes and costs roughly $0.25–1.50 per run[4] – versus tens of thousands of dollars and months of lab work for traditional experimental methods [35]. This low inference cost substantially lowers barriers to use, allowing many institutions to run these models via cloud services without owning large compute clusters.

Most scientific foundation models today are open-weight or broadly accessible[5], sometimes with restrictions due to dual-use concerns, reflecting the longstanding open-science norms of many research communities. Whether this openness will persist amid growing geopolitical competition – particularly around strategically important AI models – remains an open question.

Automating discovery: AI scientist systems

WHAT THEY ARE

AI scientist systems [36] are an emerging class of systems designed to coordinate and automate parts of the scientific research process. Their aim is to support or scale activities such as literature synthesis, hypothesis generation, experimental planning, tool selection, code execution, and evidence integration across multiple steps of inquiry.

> Google DeepMind's AI Co-Scientist [37], built on their general-purpose LLM Gemini 2.0, has a multi-agent architecture with specialised agents (Generation, Reflection, Ranking, Evolution, and Proximity agents) coordinated by a Supervisor agent. It enables "test-time compute scaling", where more inference-time computation improves output quality.
>
> ChemCrow [38], developed in an academic research setting, integrates GPT-4 with expert-designed chemistry tools, combining step-by-step reasoning with direct interaction with chemical databases and simulators.

HOW THEY ARE BUILT

The questions of data and architecture for AI scientist models are intertwined, because these systems are built on top of general-purpose, often multimodal LLMs, accessed either through proprietary APIs or as open-weight models. What distinguishes them is not so much the underlying model, but how the system is structured to perform scientific tasks.

In practice, these systems are constructed as agentic workflows in which an LLM coordinates multiple specialised sub-agents that search literature, run analyses, and test hypotheses. Rather

4. Cost estimated by multiplying typical runtime (5–30 minutes for proteins with 200–300 amino acids, per NVIDIA documentation and NIH benchmarks) by NVIDIA A100 GPU cloud pricing ($3.06/hour per AWS EC2 pricing) on the date.
5. This assessment is based on a brief survey of Epoch AI's database of notable AI models.

than merely calling external tools when convenient, the system is designed to run these agents and integrate their outputs into a shared workflow. Much like a scientist in a lab setting, the model cannot advance by reasoning alone: it must gather evidence, run analyses, and update its understanding before proceeding. This approach can be enhanced by integrating scientific tools and databases, and more advanced configurations could connect to automated laboratory equipment for physical experimentation, similar to placing the model within a well-resourced digital research facility.

AI scientist systems' architecture varies with each new model, therefore no single system is representative of the entire class. One example from 2025 is Edison Scientific's Kosmos [39]. When Kosmos is tasked with explaining a scientific phenomenon, it does not immediately generate an answer based on plausibility. Instead, it runs dedicated literature-search and data-analysis agents whose outputs populate a shared internal record. This record updates iteratively as new evidence is gathered, and the final explanation includes only claims that are linked to this evidence. If evidence is missing, the system continues searching and analysing, rather than proceeding. To date, the system's reported case studies and the composition of its development team point to a focused interest in biology and biomedicine; however, the model is being launched for general-purpose scientific use across all disciplines.

The core innovation in AI scientists therefore lies in the scaffolding: how scientific tasks are decomposed, how and when agents are triggered, how intermediate results are evaluated and accumulated, and where human oversight is required. These are scientific design choices as much as engineering ones, which means that building AI scientist systems requires close collaboration between AI developers and domain experts. Scientists are needed not just to evaluate the system outputs, but to shape how it works, from selecting tools it should use to deciding what the system should and should not do on its own.

WHAT IT TAKES TO ACCESS AND USE THEM

Because AI scientist systems are built on top of frontier LLMs – whose training costs can reach [40] tens to hundreds of millions of dollars – their development cost is often shaped by how the base model is accessed: whether it is trained in-house, used via proprietary APIs, or built on open-weight models. For example, Google DeepMind built Co-Scientist directly on their existing Gemini models, avoiding API costs entirely and making engineering and domain expert design the primary development costs.

Running AI scientist systems can be costly, with inference expenses varying widely depending on design. Open source systems such as ChemCrow typically cost around $1–10+ per query[6]. More complex proprietary systems like Kosmos run queries for many hours, which brings a high inference cost; monthly subscriptions currently cost $200[7]. The high inference cost reflects the cost of orchestrating thousands of agents in parallel: if AI scientists are understood as conductors, the price depends on how large the orchestra is. These costs are far higher than those of

6. Estimated based on ChemCrow's architecture and API pricing. ChemCrow integrates GPT-4 with 18 expert-designed chemistry tools, typically requiring 10–50 LLM API calls per task. At GPT-4 API pricing, a typical task using ~60K input and ~10K output tokens costs approximately $2.40, with the range depending on task complexity. Costs may be lower with more recent model variants.
7. Based on Edison Scientific's pricing website as of February 2026.

scientific foundation models, but they play a different role in the research pipeline. If they can shorten research timelines by weeks or months, or if their systemic use leads to scientific discoveries, their cost will be negligible compared to the cost of traditional scientific workflows.

These cost differences between systems intersect with a second, equally important issue: access, which also depends on the system and the provider. Some are open source, such as ChemCrow and Sakana AI's AI Scientist, while some remain proprietary, sometimes with limited free credits for academic use. Beyond the open-versus-closed distinction, access is tightly coupled to privacy, data governance, and compute requirements, all of which influence institutional procurement decisions. In practice, AI scientist systems require scientists to input sensitive, unpublished material such as early hypotheses, proprietary datasets, or clinical observations. Uncertainty about where this data is stored, whether it is retained, and whether it may be reused for model training can deter individual use or lead institutions to restrict adoption. Fully self-hosted deployments of open source alternatives could mitigate these concerns, but they require in-house compute capacity and at least an initial deployment effort to make systems usable at scale.

As a result, the access and the cost are intertwined with the privacy and deployment conditions of these systems. If AI scientist systems meaningfully affect scientific productivity in the long run, the ability to deploy them with a clear governance framework to clarify these conditions will become a strategic differentiator across institutions.

Do we need all three? Scientific foundation models, AI scientists, and frontier LLMs

Two high-stakes model classes now shape the strategic AI-in-science landscape, raising a practical question: **can AI scientist systems make scientific foundation models redundant, or do they remain distinct and necessary investments?** This question leads to another: **given that AI scientists are built on general-purpose LLMs, will these frontier models eventually become capable enough to render specialised AI scientists obsolete?**

Today, scientific foundation models, AI scientist systems, and frontier general-purpose models are distinctly different. But as frontier AI companies intentionally build scientific capabilities into their general-purpose models, these boundaries may blur. This dynamic means companies who develop these models gain outsized influence. Their partnership choices, open source strategy, pricing, and development priorities could quietly shape which research questions advance, which domains receive investment, and which institutions remain competitive.

This is something that science policy should prepare for. Being able to adjust procurement decisions and investment strategies as capabilities shift will be more important than getting every choice right today. If policymakers actively monitor where scientific capabilities are emerging,

build domestic capacity, and regularly reassess which systems European researchers depend on, this will help avoid locking in dependencies that are difficult to reverse.

A layered, complementary scientific AI stack

The key distinction between scientific foundation models and AI scientist systems lies in what they are designed for. Scientific foundation models are built to model the natural world itself. They are trained on specialised scientific data, and rely on architectures tailored to specific physical or biological systems. Once trained, these models tend to function as stable, reusable scientific infrastructure.

AI scientist systems operate at a different layer. Built on large language models, they are designed to organise the scientific process rather than to model natural systems directly. They coordinate tasks such as searching and synthesising literature, planning investigations, invoking tools and models, writing and executing code, and integrating evidence across sources.

These systems occupy different and complementary positions in the scientific value chain, addressing distinct bottlenecks. Scientific foundation models provide depth and reliability by learning structured representations of the natural world, enabling new forms of measurement, simulation, and prediction within specific domains. AI scientist systems, by contrast, provide breadth and coordination across the research process, linking literature, data analysis, hypothesis generation, and validation into coherent workflows that help scientists navigate and organise complex discovery processes. For example, when an AI scientist system generates a novel materials hypothesis from the literature, it can invoke a scientific foundation model to simulate the material's properties before proposing this for a physical experiment run. In this setup, the foundation model acts as the precision instrument, while the AI scientist coordinates when and how it is used, running the hypothesis validation.

Today, these systems clearly each play a unique role. Does that mean that science needs, and will continue to need, both? Scientific foundation models already have a clear footprint in science because they created capabilities that simply did not exist before. Systems like AlphaFold or GraphCast opened new doors to science, from predicting protein structures at scale to producing high-resolution weather forecasts in seconds. It is also pragmatic to keep these as standalone models, rather than folding them into a bigger AI system. A researcher who needs a protein structure prediction does not need to spin up an entire AI scientist session to get it. Compared to the value they generate, these models are light and cheap to run.

AI scientist systems, by contrast, operate in territory already occupied by human scientists, which makes their value harder to assess. Because they are built on general-purpose models, they raise concerns about hallucinations [41], which can compromise research quality or lead scientists down wrong paths. But it would be a mistake to dismiss them on that basis alone. These are not general-purpose chatbots applied to science. They are deliberately designed to act through structured workflows, updating their outputs via feedback loops that involve tools, data, and human oversight. They warrant more nuanced assessment than they currently receive.

The first generation of AI scientist systems should be read as signals of future progress. For example, in early evaluations of Kosmos, one recent AI scientist system, experts judged roughly 80%[8] of a reviewed subset of Kosmos outputs as plausible and supported by evidence. This suggests that even early systems can identify useful and sometimes novel research directions when used carefully in real workflows. However, these results also show that these systems are best utilised with human oversight, not on their own.

For that oversight to work, researchers need to be able to see how the system retrieved evidence and linked it to conclusions. Without that, unsupported claims can quietly enter research workflows. This is not hypothetical: there are already published examples [42] in machine learning research where AI-generated citations passed peer review despite being entirely fabricated. If scientists cannot properly verify what these systems produce, they will eventually stop trusting them. As *Nature Machine Intelligence* has argued [43], emergent AI scientists are a potentially transformative direction for science, but only if supported by mechanisms of traceability and reproducibility, where the AI's reasoning can be documented, referenced, and verified. This is essential if these systems are to evolve and be adopted without compromising scientific integrity.

At their current stage, AI scientist systems cannot automate end-to-end discovery cycles. However, these systems could genuinely transform how science is done, and if broader AI progress is any indication, the next generation of AI scientists will be substantially more capable. Together with scientific foundation models, they point at a layered scientific AI stack: one set of tools expanding what can be known, the other changing how knowledge is generated. These are at different levels of maturity, but together they offer European science something it has not had before.

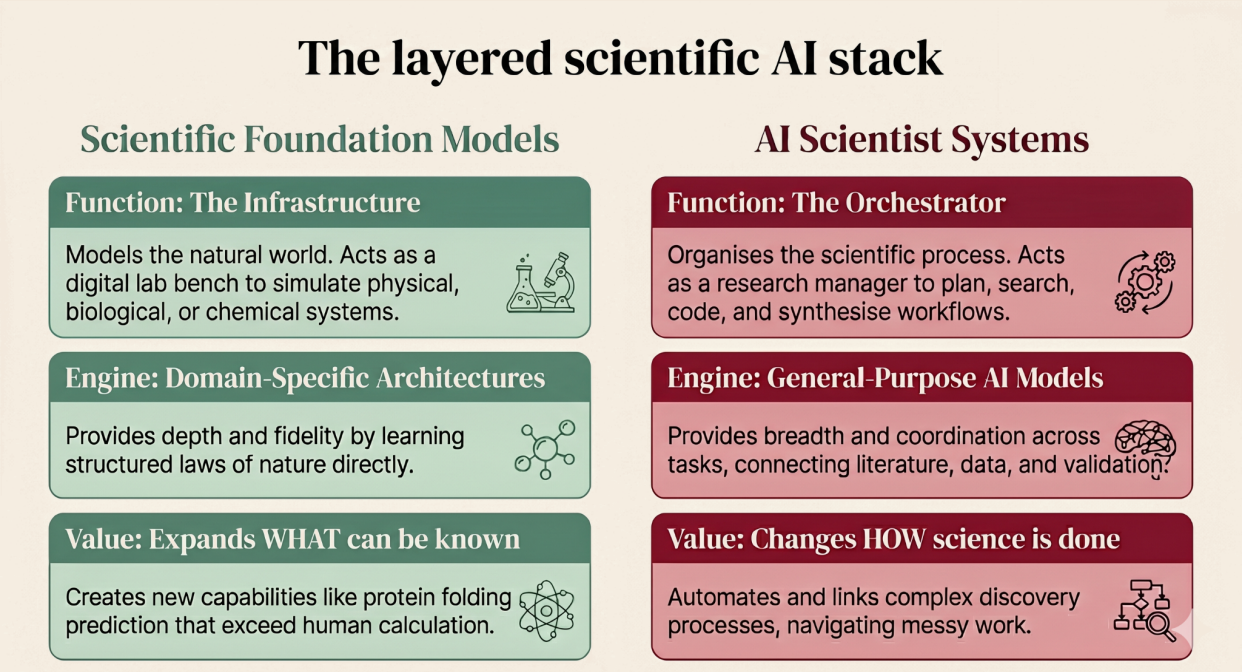

Figure2: Scientific foundation models and AI scientist systems play different but complimentary roles in the end to end scientific workflows

8. Across three biological case studies published in the study, domain experts reviewed a subset of Kosmos's generated conclusions and judged roughly 80% of them to be scientifically plausible and supported by linked literature or executed analyses.

Frontier LLMs to AI scientists: The role of tacit knowledge and scientific scaffolding

Given that current AI scientists are built on LLMs, a natural question is whether scientific capability is simply a low-hanging extension of increasingly powerful general-purpose models: will continued scaling [44] and algorithmic progress [45] cause these general-purpose models to develop scientific capabilities, or will the emerging class of AI scientists stay relevant? This intuition is appealing: as models improve in reasoning, planning, and tool use, scientific competence might seem likely to emerge as a by-product. However, most recent evidence suggests that such generalisation is unlikely without intentional design. In particular, progress in AI for science depends on addressing two key constraints: (1) systematic gaps in training data, and (2) the need for deliberate scaffolding to enable sustained, productive exploration.

First, tacit scientific knowledge exists in forms that current training approaches systematically miss. While there is no single definition, "tacit knowledge" here refers to the judgements that scientists make when formal rules fall short: how to act under uncertainty, distinguish noise from discovery, decide whether to rerun or abandon an experiment, or assess that a result is implausible despite looking correct on paper. This expertise shapes scientific outcomes, but it is difficult to explicitly record, or even encode in the first place.

Song et al. [46] provide concrete evidence of this gap. Evaluating frontier models – including GPT-5, DeepSeek-R1, Claude Sonnet 4.5, and Grok-4 – across computational scientific discovery tasks, they find shared failure modes in areas such as spectroscopic interpretation, experimental protocol design, and predicting materials' properties. These failures are consistent across models and are not primarily explained by limits in their general reasoning capabilities. Instead, they point to missing exposure to the kinds of judgement, context, and domain-specific decision-making that guide scientific work in practice.

This reflects the day-to-day, real-life practice of science. Published papers record outcomes, not the trial-and-error, troubleshooting, and informal reasoning that lead to them. As a result, models trained mainly on literature and curated datasets inherit the formal record of science, but miss a substantial portion of the practical knowledge that scientists rely on when doing research.

The same study confirms this pattern empirically. Where training data exists – for example molecular property calculations, or materials with computational validators – models perform well. Where success requires models to interpret ambiguous signals, troubleshoot based on subtle cues, or recognise contamination patterns, performance declines sharply. The bottleneck isn't reasoning capability but a dearth of examples of expert judgement in the models' training data.

Song et al. reveals a second critical finding, which further distinguishes AI scientist systems from frontier LLMs: AI scientists greatly benefit from infrastructure connecting them to scientific tools. They found that models can achieve high performances in experimental tasks if they are linked to computational validators for that task, where they can try proposals, get immediate feedback, and refine iteratively. The study calls this "serendipitous exploration" – finding solutions through structured trial-and-error rather than pure reasoning.

Scientists already use computational tools to evaluate their results extensively: chemists run validity checkers and quantum calculators, biologists query protein databases, materials researchers use energy simulators. Song et al.'s study finds that if such tools are properly connected to AI scientist systems, these systems can dramatically boost performance compared to traditional methods, essentially supercharging scientists' capacity. For example, crystal structure generation improved from 3.34% via legacy models up to 55% with LLMs and tool scaffolding, a 16× improvement.

This illustrates how the level and rigour of scaffolding largely determine performance in AI scientist systems, and where the line falls between "an LLM with many integrations" and a deliberately engineered AI scientist. At one end of the spectrum are lightweight setups, such as platforms like ToolUniverse [47], where scientists can simply integrate general-purpose language models with specific scientific tools or databases. At the other end are more fully developed AI scientist systems – such as those from Sakana, Google's Co-Scientist, or Edison Scientific's Kosmos – designed around structured workflows, iterative feedback loops, and multi-agent coordination. What belongs where on this spectrum is still evolving, and the boundaries remain fluid.

Simple integrations of LLMs into scientific tools can support experimentation, but they are unlikely to deliver transformative gains on their own. Systems designed as AI scientists differ in kind, not degree: they are built around multi-agent coordination, feedback loops, and human oversight to support end-to-end research workflows. While evidence from analyses that compare AI scientists to frontier LLMs systematically is still limited, studies like Song et al. suggest that scientific performance improves when models are deliberately scaffolded. This indicates that in the near term, AI scientist capabilities will require dedicated development.

Efforts in this direction are already underway. As mentioned earlier, companies such as OpenAI, Anthropic, and Google DeepMind now describe scientific discovery as a core use case for general-purpose models, and they are actively hiring for specialists in these domains. However, this should not be confused with product strategy. Whether AI scientist capabilities are embedded within a general-purpose model or developed as a dedicated, domain-specific system does not alter the underlying technical challenge, and frontier developers are increasingly pursuing both paths in parallel. Throughout 2025, OpenAI provided scientific reasoning capabilities through its flagship GPT models. Recently in 2026, they released a reasoning model purpose-built for biology, drug discovery, and translational medicine called GPT-Rosalind [48], currently restricted to a trusted-access programme due to dual-use safety considerations. Google has been on this path for longer, launching a dedicated Co-Scientist as a distinct system layered on Gemini in early 2025: it is currently in beta phase. Regardless of the launch strategy, what developers concretely need for building these capabilities is the same: deliberate system design, model scaffolding, targeted data, and infrastructure integration to make scientific outputs reliable.

Current evidence suggests that frontier models will not simply become AI scientists through scaling alone. Progress in scientific capability depends on building infrastructure that enables iterative validation, structured tool use, and domain-specific judgement. Larger and more capable models may have better performance, but these system-level components remain necessary. The central question for advancing scientific capabilities is where the main bottlenecks lie, from

specific datasets to workflow design, and how to identify them in order to direct investment effectively.

Strategic coordination: Matching investment to the distinct data needs of each layer

While scientific foundation models and AI scientist systems are distinct layers of a scientific AI stack, a shared bottleneck shaping their progress is data. However, the two paradigms require fundamentally different types of data and data strategies.

For scientific foundation models, the data challenge is primarily one of scale and coverage. The type of large-volume scientific data required for scientific foundational models is typically produced by national laboratories, observatories, and major research institutes. Expanding, standardising, and maintaining these data pipelines will require coordinated action. AI scientist systems, by contrast, are constrained by a different absence: missing data about how science actually happens. The reasoning, iteration, and judgement guiding discovery exists mostly as tacit knowledge – things like experimental troubleshooting, informal decisions, and interpretation under uncertainty. This knowledge is passed on through on-the-job learning but rarely documented as data. As the Institute for Progress has argued [49], this type of data requires capturing reasoning patterns, workflows, iterative trial-and-error, and both successes and failures in research, none of which have been typically represented in the datasets used for AI training. They propose creating "unstructured data generation labs" that would comprehensively record everything about how scientists do science: bodycam footage, keystroke logs, experimental protocols, and failed attempts. This approach raises concerns about surveillance and privacy, and more targeted methods might prove more efficient. But the core insight holds: AI systems need systematic data on scientific processes, and we lack the frameworks and infrastructure to capture this data at scale.

Scientific foundation models and AI scientist systems require different kinds of infrastructure. Without clearly distinguishing which layer is being targeted, even well-funded initiatives risk missing the bottlenecks that actually constrain progress. The same logic applies one level deeper: investors should ask themselves not only whether AI capabilities can resolve a binding scientific constraint, but also which capability specifically.

For example, clinical-trial timelines are a bottleneck [50] in cancer research. One way of addressing this bottleneck is regulatory or procedural, where AI plays no role; another potential path is developing [51] AI-driven digital twins for trial simulation and coordinating the patient data needed to make them reliable. This level of problem identification is what reveals whether and which AI capability is needed, and helps direct effort toward the bottlenecks AI is best suited to resolve.

Current public initiatives on AI for science show both promise and clear limits along these lines. Both the EU's RAISE and the U.S. Genesis Mission place strong emphasis on coordinating data and resources across public and private actors, within their respective scopes and mandates. This focus on coordination is valuable - but it becomes far more effective when it is informed by how

science is actually practiced: where models succeed or fail, which workflows benefit from automation, and what kinds of data are genuinely missing. Scientists therefore play a critical role, not just as users, but as evaluators whose feedback can help steer coordination toward real needs rather than assumed ones.

RAISE's explicit intent to mobilise the private sector and global partners – including philanthropic organisations – through a planned pledging exercise is particularly valuable. Scientific philanthropy is increasingly recognising that funding shared data and workflow infrastructure can be as impactful as funding individual labs. For example, Renaissance Philanthropy's AI for Science Dataset call [52], launched with the UK Department for Science, Innovation & Technology, targets datasets that could become foundational infrastructure in domains that are UK science policy priorities, from fusion energy to medical research. These areas align closely with RAISE's focal domains, creating opportunities for shared learning and complementary investment. Meanwhile, the UK ARIA's new AI Scientist programme [53] experiments with what scientific processes AI can help automate, directly informing where workflow infrastructure gaps exist. The private sector can play a similar role: DeepMind's recent proposal [54] on AI data stocktakes for fusion energy used structured expert consultation to map a field's data gaps and translate them into fundable projects.

Together, these initiatives address the two distinct layers of AI for science: foundation models that require large, standardised datasets, and AI scientist systems that depend on scientific workflow learnings. While private companies and philanthropy can accelerate progress, they are less likely to sustain easily accessible, interoperable infrastructure for scientists over the long term. Public investment can be better positioned to coordinate standards and fund shared data and workflow systems as scientific infrastructure. Without this systemic investment, AI for science will advance unevenly, with capability concentrated in a few institutions rather than embedded across the broader research system.

Key policy takeaways

"AI for science" is not a single category, and different aspects require different investment strategies

"AI for science" covers at least two strategically distinct classes of AI: scientific foundation models and AI scientist systems. Scientific foundation models create entirely new scientific tools – systems that can simulate proteins, climate systems, or materials with a level of fidelity that did not previously exist. AI scientist systems, on the other hand, accelerate scientific progress itself by running many structured lines of inquiry in parallel, learning from each iteration and feeding that learning back into the next. While complementary, they depend on different forms of data, expertise, and technical infrastructure, and they contribute to scientific progress in distinct ways. For decision-makers, this means they cannot be treated as interchangeable or evaluated through a single lens: each occupies a different place in the scientific value chain and requires tailored assessment and investment strategies.

The AI for science roadmap should be designed for a fast-moving frontier, not a fixed alloca - tion

What "AI for science" even encompasses has changed in just a couple of years. Not long ago it mostly meant large models specialised in a single task, such as protein structure prediction; now it also covers systems built for end-to-end scientific discovery workflows, and the line between "scientific AI" and "frontier AI" will keep moving. Funding decisions locked into one paradigm would already have missed an entire layer that has since emerged. The right response is not to predict the next shift but to stay able to act on it: testing and funding existing tools to learn what actually works, backing early bets and scaling the ones that pay off, and winding down what stalls. In practice, this means AI for science roadmaps should be updated on an ongoing basis rather than set once, with funders willing to move faster than a standard multi-year programme allows.

AI scientist systems are early-stage, but have transformative potential

Evidence on how robust AI scientists are across disciplines is still limited. Many current systems are built on general-purpose language models and inherit their constraints, including reliability issues. Even so, these systems represent one of the first attempts to restructure how research is conducted, not just through automating individual tasks, but organising scalable, repeatable scientific workflows. If they mature, they could shorten experimental cycles, run parallel lines of inquiry, and increase research productivity in institutions that integrate them effectively. Since many are skeptical about AI scientists, there's a risk that we will fail to anticipate how quickly these capabilities can mature. A measured policy stance here is to support experimentation and development of AI scientist systems with scientists firmly in the loop, acknowledging that failures are valuable learnings that can help us improve AI's science capabilities. This support should extend beyond funding: frameworks on intellectual property, data governance, and research integrity must be updated in parallel.

Current frontier LLMs are unlikely to act as AI scientists, but this could change

Today's general-purpose frontier models exhibit systematic failures on scientific tasks, suggesting that scaling alone might not deliver reliable, generalisable scientific capability. General-purpose models lack targeted data, validation, and tool integration. AI scientist systems respond to this gap through deliberate scaffolding: structured workflows, external tools, and feedback loops tailored to scientific use. However, with sufficient engineering effort to build the right model infrastructure, and targeted data, frontier model developers may be able to incorporate these capabilities into general-purpose models over time. There is evidence that this is a priority for leading frontier AI developers.

Data is the central constraint for AI for science – but different layers need different types of data

Scientific foundation models rely on large-scale, domain-specific data, mostly acquired through labs and observatories. This makes them a natural focus for coordination-heavy public initiatives such as the EU's RAISE programme or the U.S. Genesis Mission. These efforts are important and necessary, but they address only part of the data challenge. AI scientist systems are constrained by a different gap: the absence of data about how science is actually done – workflow practices,

troubleshooting, failed experiments, and expert judgement. To address this, one organisation has proposed [49] setting up laboratories where end-to-end research processes are recorded (compare embodied robotic systems, which are trained using hours of video of people performing household tasks). . Recording research this way raises legitimate concerns about norms around privacy and credit. But if such efforts lead to a differential leap in scientific productivity, this is a bottleneck that investment and coordination initiatives should aim to address deliberately.

Scientists must act as evaluators, not just end users Scientists are essential for guiding strategic AI-in-science investment. As AI capabilities evolve rapidly, hands-on experience and structured feedback from scientists – on where models fail, which workflows benefit from automation, what data is missing, and what reliability means in practice – are critical. At the same time, scientists operate under strong incentives to preserve rigour, reproducibility, and research quality, which can make experimentation with immature tools risky or costly. With many researchers already overstretched, dedicated mechanisms and intermediaries are needed to lower the cost of safe experimentation and shared learning. Organisations such as SCIANCE [55] can play a key role in bridging scientific practice and EU AI-in-science policy by enabling faster, more systematic feedback loops.

Investment should target capability gaps, not only scale

Given the vast scope of AI for science, effective investments in the field could look less like traditional grants and more like targeted capability-building: specific datasets that fill a missing gap, workflows and infrastructure, and even negative results that generate learnings can be meaningful, which rarely fit standard publication-based metrics of scientific excellence. A recent U.S. federal medical research funding initiative [56] reflects this by adopting a "scouting" approach and enabling re-granting mechanisms, recognising that the government cannot always identify the most critical gaps. Intermediaries such as philanthropic initiatives can move faster, take risks, and target under-funded capabilities, translating coordination into real performance gains across the scientific AI stack.

Access and governance will shape who benefits, not just model quality.

Access and governance, and not just how good models are, will determine who benefits from AI for science. Science has traditionally been built on openness – shared data, methods, and results – but as AI becomes more capable and more strategic, this openness may come under pressure: the most capable scientific AI systems released in 2025 and 2026 launched [57] with restricted, vetted access rather than broad availability, often justified by dual-use safety considerations. Whether AI-enabled scientific capability spreads widely or concentrates in a few places will depend on the terms of public–private partnerships, access frameworks, and interoperability choices, such as lightweight tool-based approaches versus expensive, managed systems. Policy decisions in these areas can quietly create long-lasting advantages for some research ecosystems over others. In a changing geopolitical context, maintaining broad access while developing reliable, sovereign AI-in-science capabilities may become an increasingly important strategic consideration.

Conclusion

AI has the potential to reshape how scientific knowledge is produced, not only by accelerating analysis but by changing how hypotheses are generated, experiments are designed, and results are validated. To respond effectively, AI for science should be understood as a layered capability rather than a single technology. One layer consists of scientific foundation models – large AI systems trained on extensive scientific datasets that can model natural phenomena across domains. The second layer consists of AI scientist systems that orchestrate the research process itself, from generating hypotheses to designing and validating experiments.

As funding under RAISE and Framework Programme 10 takes shape, Europe's public investments should treat AI for science as a stack of scientific infrastructure, with distinct and targeted initiatives to address the data and platform needs of each layer. Public funding should build both layers of this stack: expanding the large scientific datasets needed for foundation models while also supporting the experimental platforms and workflow infrastructure required for AI scientist systems. Developers and evaluators should centre real research practice: working scientists should be directly involved in testing and iteration. Access rules, data standards, and governance frameworks should likewise be designed with long-term capability in mind, as they will shape which institutions remain able to develop and apply these systems.

If these foundations are not built early, AI-enabled discovery may concentrate in a small number of companies and countries that control the necessary data, infrastructure, and experimentation platforms. This applies to Europe very concretely. Despite its world-leading research base, it lags behind the US and China in homegrown AI models and compute resources that increasingly underpin AI-driven science. Matching investment to the distinct needs of each layer of the scientific AI stack should therefore be core to Europe's AI for science strategy — and to its ambition of staying at the forefront of global scientific progress.

Reference List

[1] European Commission, European AI for science Strategy, Research and Innovation, October 8, 2025, https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/our-digital-future/european-ai-science-strategy_en

[2] U.S. Department of Energy. Energy Department Launches 'Genesis Mission' to Transform American Science and Innovation Through the AI Computing Revolution. U.S. Department of Energy, November 24, 2025. https://www.energy.gov/articles/energy-department-launches-genesis-mission-transform-american-science-and-innovation

[3] Soete, L., Verspagen, B., and Ziesemer, T.H.W., Economic Impact of Public R&D: An International Perspective, Industrial and Corporate Change, vol. 31(1), 2022, https://academic.oup.com/icc/article/31/1/1/6377306

[4] Jones, B.F., and Summers, L.H., A Calculation of the Social Returns to Innovation, NBER Working Paper No. 27863, 2020, https://www.nber.org/papers/w27863

[5] Aristodemou, L., Appelt, S., van Beuzekom, B., and Galindo-Rueda, F., Assessing the Relevance of R&D Funding Towards Societal Goals, OECD Science, Technology and Industry Working Papers No. 2025/25, 2025, https://

www.oecd.org/en/publications/assessing-the-relevance-of-r-d-funding-towards-societal-goals_bafcdc7b-en.html

[6] Frontier Economics. Rate of Return to Investment in R&D. Department for Science, Innovation and Technology, 2023. https://www.frontier-economics.com/media/015adtpq/rate-of-return.pdf

[7] Hassabis, D., Jumper, J., Kohli, P., & Koivuniemi, A., on behalf of the AlphaFold Team. (2025, November 25). AlphaFold: Five years of impact. Science, https://deepmind.google/blog/alphafold-five-years-of-impact/

[8]European Commission. (2025). CORDIS Results Pack on AI in life sciences: Harnessing the transformative power of AI for scientific discovery (July 2025) [PDF]. Publications Office of the European Union.

https://publications.europa.eu/resource/cellar/c4d2180b-5aec-11f0-a9d0-01aa75ed71a1.0001.03/DOC_1

[9] Fudan University and Shanghai Academy of AI for Science, AI for Science 2025, Nature (Advertisement Feature), 2025, https://www.nature.com/articles/d42473-025-00161-3

[10] Berman, B., Chubb, J., and Williams, K., The Use of Artificial Intelligence in Science, Technology, Engineering, and Medicine, The Royal Society, 2024, https://royalsociety.org/-/media/policy/projects/science-in-the-age-of-ai/science-ai-taxonomy-report.pdf

[11] The Royal Society. Science in the Age of AI: How Artificial Intelligence Is Changing the Nature and Method of Scientific Research. The Royal Society, 2024. https://royalsociety.org/-/media/policy/projects/science-in-the-age-of-ai/science-in-the-age-of-ai-report.pdf

[12] Wang, Hanchen, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, et al. "Scientific Discovery in the Age of Artificial Intelligence." *Nature Reviews Methods Primers* 3 (2023). https://doi.org/10.1038/s41586-023-06221-2

[13] Chen, Q., Yang, M., Qin, L., Liu, J., Yan, Z., Guan, J., Peng, D., Ji, Y., Li, H., Hu, M., Zhang, Y., Liang, Y., Zhou, Y., Wang, J., Chen, Z., and Che, W., AI4Research: A Survey of Artificial Intelligence for Scientific Research, arXiv:2507.01903, 2025, https://arxiv.org/abs/2507.01903

[14] Gridach, M., Nanavati, J., Zine El Abidine, K., Mendes, L., and Mack, C., Agentic AI for Scientific Discovery: A Survey of Progress, Challenges, and Future Directions, arXiv:2503.08979, 2025, https://arxiv.org/abs/2503.08979

[15] Hu, M., et al., A Survey of Scientific Large Language Models: From Data Foundations to Agent Frontiers, arXiv:2508.21148, 2025, https://arxiv.org/abs/2508.21148

[16] Rodriques, S., and Hinks, M., 'Kosmos: An AI Scientist for Autonomous Discovery', Edison Scientific Blog, November 5, 2025, https://edisonscientific.com/articles/announcing-kosmos

[17] European Commission. (2025, October 8). *European AI for science Strategy*. Research and Innovation — European Commission. https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/our-digital-future/european-ai-science-strategy_en

[18] U.S. Department of Energy. (2025, November 24). *Energy Department launches 'Genesis Mission' to transform American science and innovation through the AI computing revolution*. https://www.energy.gov/articles/energy-department-launches-genesis-mission-transform-american-science-and-innovation

[19] Girishankar, N., and Borges, C., The Genesis Mission: Can the United States' Bet on AI Revitalize U.S. Science?, CSIS Critical Questions, December 4, 2025, https://www.csis.org/analysis/genesis-mission-can-united-states-bet-ai-revitalize-us-science

[20] OpenAI, OpenAI for Science, https://openai.com/science/, (accessed February 6, 2026)

[21] Anthropic, 'Introducing Anthropic's AI for Science Program', Anthropic News, May 5, 2025, https://www.anthropic.com/news/ai-for-science-program

[22] Gottweis, J., and Natarajan, V., 'Accelerating Scientific Breakthroughs with an AI Co-Scientist', Google Research Blog, February 19, 2025, https://research.google/blog/accelerating-scientific-breakthroughs-with-an-ai-co-scientist/

[23] FutureHouse, Automating Scientific Discovery, https://www.futurehouse.org/, (accessed February 6, 2026)

[24] Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., et al., On the Opportunities and Risks of Foundation Models, Stanford Center for Research on Foundation Models (CRFM), 2021, https://crfm.stanford.edu/report.html

[25] Bonev, B., Kurth, T., Koch, M., Foster, D., Robinson, N., Paris, A., Pritchard, M., and Keller, A., 'FourCastNet 3 Enables Fast and Accurate Large Ensemble Weather Forecasting with Scalable Geometric ML', NVIDIA Developer Blog, July 29, 2025, https://developer.nvidia.com/blog/fourcastnet-3-enables-fast-and-accurate-large-ensemble-weather-forecasting-with-scalable-geometric-ml/

[26] Lam, R., GraphCast: AI Model for Faster and More Accurate Global Weather Forecasting, Google DeepMind Blog, November 14, 2023, https://deepmind.google/blog/graphcast-ai-model-for-faster-and-more-accurate-global-weather-forecasting

[27] Djouama, I., Nabil, K., and Seghir, R., ML-Based Weather Forecasting Models: A Comparative Study, Lecture Notes in Networks and Systems, vol. 1170, Springer, 2024, https://link.springer.com/chapter/10.1007/978-3-031-73344-4_33

[28] "AlphaFold," Wikipedia, https://en.wikipedia.org/wiki/AlphaFold (accessed February 6, 2026)

[29] Robertson, B.E., Tacchella, S., Johnson, B.D., et al., Morpheus Reveals Distant Disk Galaxy Morphologies with JWST, The Astrophysical Journal Letters, 942(2), L42, 2023, https://iopscience.iop.org/article/10.3847/2041-8213/aca086

[30] "James Webb Space Telescope," Wikipedia, https://en.wikipedia.org/wiki/James_Webb_Space_Telescope (accessed February 6, 2026)

[31] Johns, A., Seaton, B., Gal-Edd, J., Jones, R., Fatig, C., and Wasiak, F., James Webb Space Telescope -- L2 Communications for Science Data Processing, Proceedings of SPIE 7016, 2008, https://ntrs.nasa.gov/api/citations/20080030196/downloads/20080030196.pdf

[32] Reiser, P., Neubert, M., Eberhard, A., et al., Graph Neural Networks for Materials Science and Chemistry, Communications Materials, 3, 93, 2022, https://www.nature.com/articles/s43246-022-00315-6

[33] Subramanian, S., Harrington, P., Keutzer, K., Bhimji, W., Morozov, D., Mahoney, M.W., and Gholami, A., Towards Foundation Models for Scientific Machine Learning: Characterizing Scaling and Transfer Behavior, NeurIPS, 2023, https://proceedings.neurips.cc/paper_files/paper/2023/file/e15790966a4a9d85d688635c88ee6d8a-Paper-Conference.pdf

[34] European High Performance Computing Joint Undertaking (EuroHPC JU), Our Supercomputers, https://www.eurohpc-ju.europa.eu/supercomputers/our-supercomputers_en, (accessed February 6, 2026)

[35] "Rosetta@home," Wikipedia, https://en.wikipedia.org/wiki/Rosetta%40home (accessed February 6, 2026)

[36] Lu, C., Lu, C., Lange, R.T., Foerster, J., Clune, J., and Ha, D., The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery, arXiv:2408.06292, 2024, https://arxiv.org/abs/2408.06292

[37] Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., ... Natarajan, V. (2025). *Towards an AI co-scientist*. arXiv. https://arxiv.org/abs/2502.18864

[38] Bran, A. M., Cox, S., Schilter, O., Baldassari, C., White, A. D., & Schwaller, P. (2023). *ChemCrow: Augmenting large-language models with chemistry tools* (arXiv:2304.05376) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2304.05376

[39] Mitchener, L., Yiu, A., Chang, B., et al. (2025). *Kosmos: An AI Scientist for Autonomous Discovery* (arXiv:2511.02824). https://arxiv.org/abs/2511.02824

[40] Buchholz, K. (2023). *The Surging Cost of Training AI Models*. Visual Capitalist. https://www.visualcapitalist.com/the-surging-cost-of-training-ai-models/

[41] Emslie, K. (May 23, 2024). *LLM Hallucinations: A Bug or A Feature?* Communications of the ACM. https://cacm.acm.org/news/llm-hallucinations-a-bug-or-a-feature/

[42] Ansari, S. (2026). *Compound Deception in Elite Peer Review: A Failure Mode Taxonomy of 100 Fabricated Citations at NeurIPS 2025* (arXiv:2602.05930) [cs.DL]. https://arxiv.org/abs/2602.05930

[43] Nature Machine Intelligence. (2026). *Multi-agent AI systems need transparency*. https://www.nature.com/articles/s42256-026-01183-2

[44] Samborska, V. (January 20, 2025). *Scaling up: how increasing inputs has made artificial intelligence more capable*. Our World in Data. https://ourworldindata.org/scaling-up-ai

[45] Epoch AI. (March 12, 2024). *Algorithmic progress in language models*. Epoch AI Blog. https://epoch.ai/blog/algorithmic-progress-in-language-models

[46] Song, Z., Lu, J., Du, Y., Yu, B., Pruyn, T. M., Huang, Y., et al. (2025). *Evaluating large language models in scientific discovery* (arXiv:2512.15567) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2512.15567

[47] ToolUniverse. *700+ Scientific Tools to Democratize AI Scientists*. https://aiscientist.tools/ (accessed February 6, 2026).

[48] OpenAI. (April 17, 2026). *Introducing GPT-Rosalind*. https://openai.com/index/introducing-gpt-rosalind/

[49] Reinhardt, B. (August 13, 2025). *Teaching AI How Science Actually Works*. Institute for Progress. https://ifp.org/teaching-ai-how-science-actually-works/

[50] Javorsky, E. (2026). *How AI Can, and Can't, Cure Cancer*. Future of Life Institute. https://curecancer.ai/

[51] Thangaraj, P.M., Shankar, S.V., Huang, S., Nadkarni, G.N., Mortazavi, B.J., Oikonomou, E.K., and Khera, R. (2026). A novel digital twin strategy to examine the implications of randomized clinical trials for real-world populations. *npj Digital Medicine*, 9, 329. https://doi.org/10.1038/s41746-026-02464-1

[52] Renaissance Philanthropy. (2025). *AI for Science Datasets: Request For Proposals*. https://www.renaissancephilanthropy.org/ai-for-science-dataset-rfp

[53] ARIA. (2026). *AI Scientist*. https://www.aria.org.uk/ai-scientist/

[54] Griffin, C., Wallace, D., and Brown, T. (April 30, 2026). *Science Needs AI Data Stocktakes: A Proof-of-Concept for Fusion Energy*. Google DeepMind. https://deepmind.google/public-policy/science-needs-ai-data-stocktakes/

[55] SCIANCE. *Resource for AI Science in Europe (RAISE)*. https://www.sciance.eu/ (accessed February 6, 2026).

[56] Office of Representative Josh Harder. (2026). *NIH: Harder Unveils Landmark Legislation to Supercharge Medical Breakthroughs at Top Science Agency*. U.S. House of Representatives. https://harder.house.gov/media/press-releases/nih-harder-unveils-landmark-legislation-to-supercharge-medical-breakthroughs-at-top-science-agency

[57] Perrigo, B. (April 24, 2026). *'Too Dangerous to Release' Is Becoming AI's New Normal*. TIME. https://time.com/article/2026/04/24/claude-mythos-chatgpt-rosalind-release-dangerous/

About the author

Bengüsu Özcan

Senior Researcher, Frontier AI Governance, Arq Foundation

Bengüsu's work at Arq spans international AI governance coordination, scenario planning, and stakeholder engagement across European and global audiences. Previously, at the Centre for Future Generations, she co-authored *Advanced AI: Possible Futures* and *Europe and the Geopolitics of AGI*. She co-founded AI Safety Türkiye and is a published novelist. She holds an MA in Quantitative Methods in Social Sciences from Columbia University and a BSc in Industrial Engineering and Psychology from Sabancı University.

Arq

This report is part of Arq's research. Arq is a Brussels-based think tank focused on preparing Europe and the world for transformative AI. For more, or to get in touch, visit **arq.foundation**.