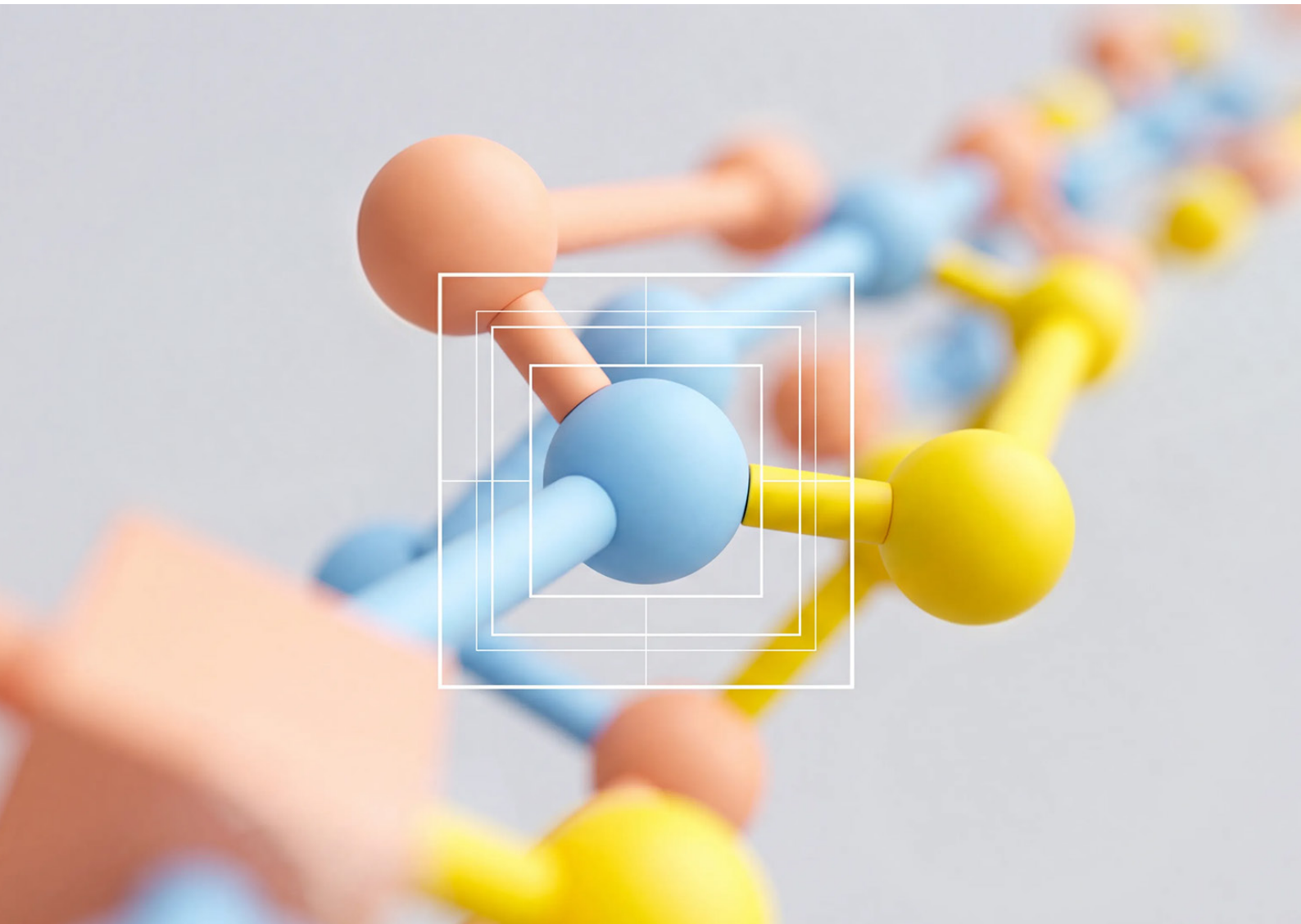
July 2026

Conjecture Machines: AI agents and the new validation bottleneck in science

By Don Wallace, Conor Griffin, Sean O'Neill, Thang Luong, Owen Larter



Over the past year, agents have transformed what it means to code. Scientific research may be next — from proposing novel hypotheses, to designing experiments, to discovering algorithms that improve on the best that humans have designed.

This raises urgent questions for policymakers and science funders, the biggest being how to validate the coming wave of AI-generated ideas.

New approaches are also needed to ensure that scientists can access agents, that datasets are agent-ready, and that peer review is kept afloat.

We sat down with 10 researchers and engineers from Google DeepMind to try to figure out what comes next.

Introduction

Microbiologist José Penadés and his team at Imperial College London took most of a decade to work out how a family of superbugs spreads antibiotic resistance. The result was unpublished, known only inside his lab. Then, in 2024, he described the problem to Co-Scientist, an AI agent from Google DeepMind.

Within two days, Co-Scientist returned five potential explanations, ranked in priority. Number 1 was the same hypothesis Penadés's team had spent so long working to prove: that some superbugs acquire tails from viruses and use them as "keys" to jump between host species. Stunned, his first thought was that his computer had been compromised, so he emailed Google to check. Confirmed: no peeks taken. If he could have gone back in time, that insight would have saved his team years.

An AI agent such as Co-Scientist is a large language model (LLM)-based system given a goal and the tools to pursue it. Unlike a query-answering chatbot, an agent can plan how to achieve the goal you give it, breaking it down into steps, running multiple subagents and processes in parallel, and detecting and correcting errors as it goes. It can also engage with the wider world, for example by calling online databases and tools, or writing code to operate robots.

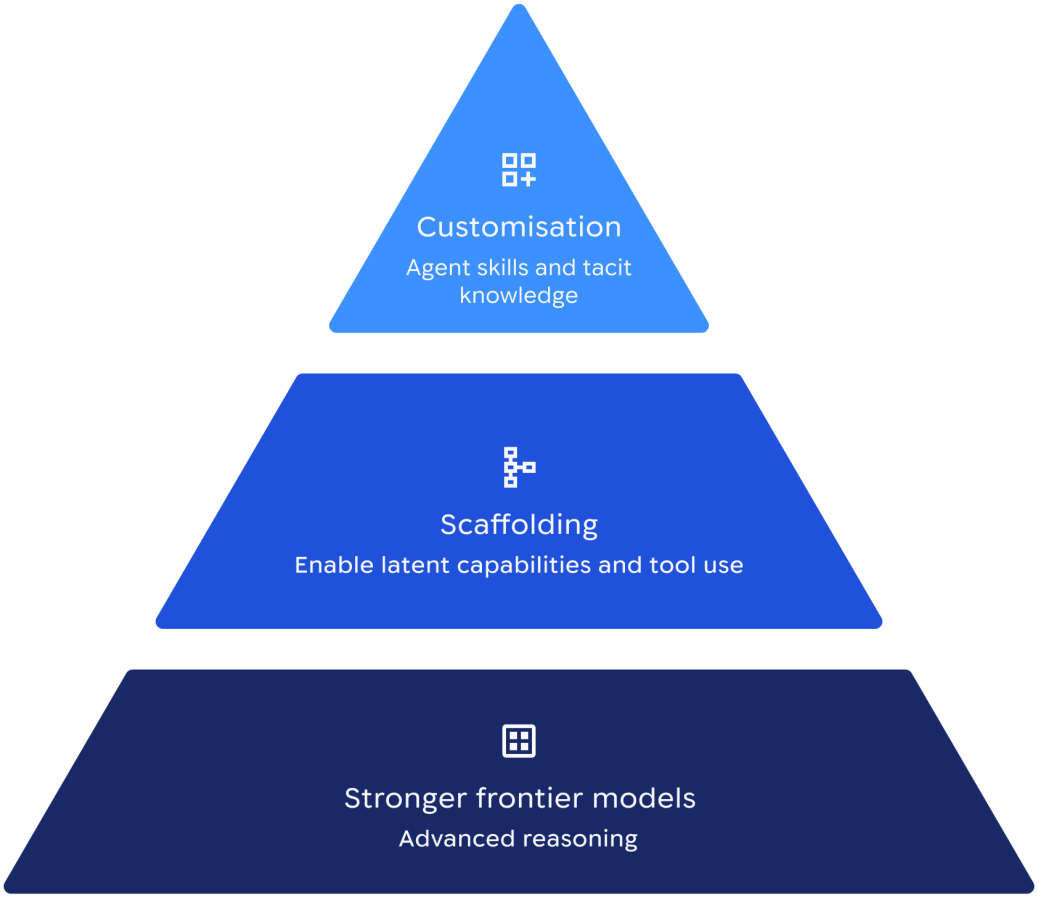
The arrival of AI agents is timely. Researchers face a rapidly growing "burden of knowledge", with more to learn before they can meaningfully contribute. The questions that count in drug discovery, climate modelling, materials design, and biology are becoming too complex and interdisciplinary for human teams to tackle at pace.

Science depends in part on a social infrastructure that has gradually evolved: labs, teams, institutions, peer review, grant funding, and networks that support the accumulation of shared knowledge. The era of AI agents will challenge that infrastructure, and at times demand rapid change — some of which is perhaps overdue.

Software engineering has been going through a version of this. In barely a year, coding agents reshaped the working practices of engineers. AI agents are now starting to rreshape science. Science is messier, but it also suits agents in certain ways: a large literature available as text, huge databases, and some workflows already built around code. Science also has an existing generation of specialist AI models that agents can put to work, from protein and materials design tools to state-of-the-art weather forecasting.

Why are AI agents suddenly so useful?

Anybody who has worked with early iterations of AI agents may be sceptical of their utility in science, where rigour and reliability are essential. But three forces are making today's agents more useful: stronger frontier models, "scaffolding" that enables agents to elicit greater capabilities from those frontier models, and customisability that allows scientists to tailor agents to their needs.



Stronger frontier models. Leading models now outpace human experts on demanding scientific knowledge benchmarks, such as Humanity's Last Exam and FrontierMath. Perhaps more significant is their new depth of thinking, driven by inference-time reasoning, a technique in which models work through problems in extended steps, exploring and revising before committing to an answer.

In mathematics and computer science, where breakthroughs can rest on reasoning alone, models are producing impressive results. But sharper reasoning helps beyond mathematics, enabling agents to draw on the "long tail" of findings that are often buried in papers, make scientific connections across fields, judge which tools to call and when, and catch their own errors as they go.

Scaffolding. This is a type of "harness" that enables AI agents to better elicit the latent capabilities that reside in base models but which do not emerge by default. This code "wraps" a frontier model, giving it structure, memory, and the ability to access and use tools including code execution, scientific software, and literature search. It also lets agents call on specialised AI models and interact with other agents.

Early agent scaffolds were bespoke: they gave a particular agent detailed instructions for how to plan, carry out tasks, and use tools. But often they did not transfer well to other agent systems, or even to later generations of the frontier models they depended on. New standards, such as protocols for how agents can communicate with each other, should reduce the need for so much custom scaffolding.

Customisation. Agents can be tailored to a discipline, a lab's workflow, or an individual researcher. A key mechanism for that is the current proliferation of agent "skills" that users can create to their own detailed specs.

Base models hold a great deal of explicit scientific knowledge, the kind written down in papers and textbooks. What they lack is tacit knowledge: the hard-to-articulate craft, built up over years of practice and failure, that lets a scientist coax a cell line into growing or get a simulation to run well. Until now, scientists using AI tools had to convey their methodological know-how, processes, and preferences the hard way: through detailed prompts, bespoke scaffolding, or model retraining.

Agent skills make some of this know-how portable. They are instruction sets, often just simple text files, that tell an agent how to perform a task and what to produce. Skills are relatively easy to produce, share and accumulate. As a scientist uses their agent more, the agent can also draw on these interactions, making the experience more personalised. Scientific labs and institutions can also make their proprietary data — such as old lab notebooks — securely available to their agents.

Natasha Latysheva, a computational biologist, says she has distilled aspects of her research process into skills. "While AI agents aren't fully reliable yet, I think the trend is clear. Scientific research will shift from hands-on execution to high-level orchestration," she says. "We'll start our workday by reviewing the experiments and analyses our agents ran overnight, tweaking their direction and guiding their attention."

> "While AI agents aren't fully reliable yet, I think the trend is clear.

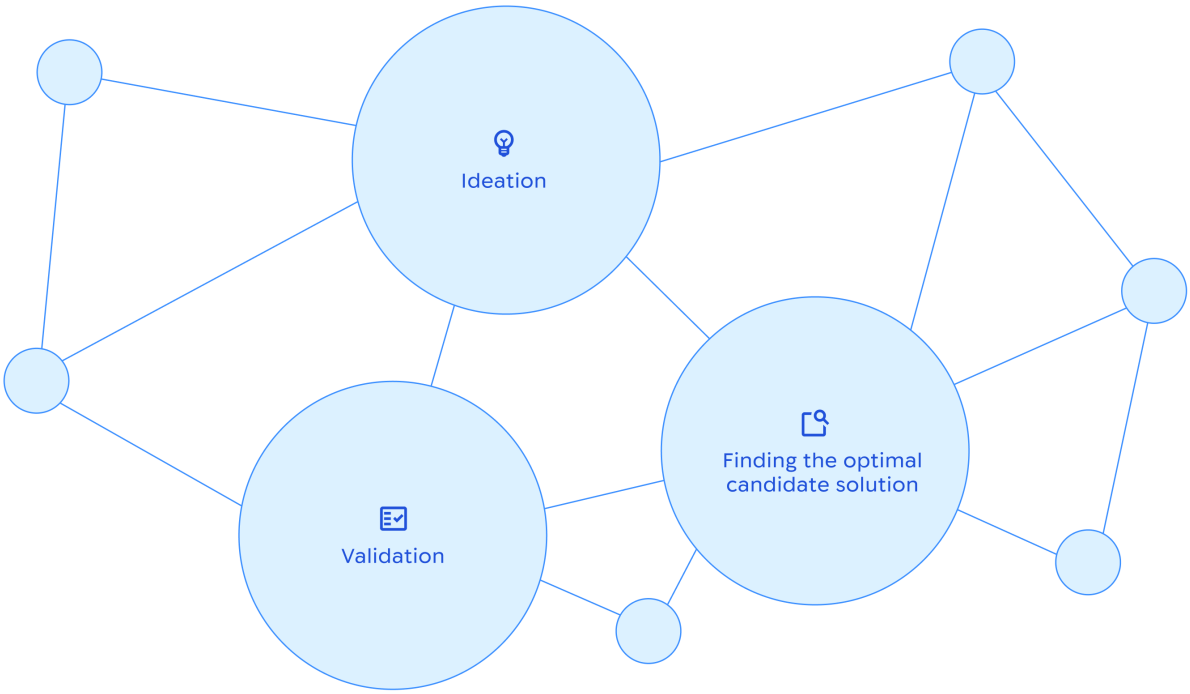
How AI agents are changing (and not changing) science

The most immediate change for scientists is also the most mundane: the availability of smart, tireless digital assistants. Researchers are handing off more of their daily grind — sifting the literature, querying databases, orchestrating analyses, curating data, writing grant proposals — to agents that do it in minutes rather than hours or days, then report back.

Agents also open up new possibilities. "Researchers can suddenly do things they couldn't do at all before," says Matej Balog, a Senior Staff Research Scientist. For example, plenty of scientists lack the skills to build the software tools and pipelines they need. This is partly because developing software for science is hard. But also because most scientists are not deeply trained in programming, and science has not traditionally provided competitive career paths for dedicated research software engineers in academic labs.

Agents are particularly strong at writing code because it is a domain with an enormous amount of training data, where correctness can typically be checked automatically. A scientist can now describe what they need in natural language — "write me a script to clean and merge these three datasets", "build me an interactive browser-based tool to explore this output" — and get serviceable code in minutes.

But ultimately, the biggest change for scientists is a structural one. Agents are making it easier to come up with ideas and proposed solutions for problems they are working on, but are not yet providing the same uplift when it comes to validating them.



Ideation

Most scientists have no shortage of ideas. The challenge is knowing which ones to pursue. This is where agents are starting to help. An agent can digest a field's accessible literature, making connections across disciplines that no single researcher would have time to trace. Existing Deep Research tools already use subagents to do this. But agents optimised for ideation go even further. Co-Scientist, for example, directs a raft of subagents to generate diverse hypotheses, critique them, rank them, and iterate — mimicking aspects of how a human research group operates, but at remarkable speed.

Gary Peltz at Stanford University used the tool in his hunt for existing drugs that could be repurposed to treat liver fibrosis, the scarring process behind 1.4 million cirrhosis deaths a year. Based on his own literature review and decades of expertise, he picked two candidate drugs. Co-Scientist picked three. Neither of Peltz's picks showed any benefit in assays with live human liver cells. Two of Co-Scientist's picks not only blocked fibrosis but also promoted liver cell regeneration.

This is a positive example, but any LLM-based system is fallible — even the strongest reasoners can still make things up. "A single hallucinated claim on page 10 of an output can invalidate the whole thing," says Vivek Natarajan, a Co-Scientist lead. Missing such errors wastes time and resources, and catching them can require a scientist with deep expertise. This fallibility makes scientists understandably cautious about AI agents, and reducing it is a top priority for those building these systems. To function effectively in a research setting, agents cannot act as black boxes that simply output answers; they must expose their reasoning and serve as transparent collaborators.

> "A single hallucinated claim on page 10 of an output can invalidate the whole thing.

One key challenge is imbuing agents with what Natarajan calls "epistemic humility": models that know when they don't know, and say so. He points to AlphaFold, the protein-structure predictor, which is prized partly because it reports how confident it is in each aspect of each prediction, so researchers know when to trust it and when to reach for other methods. Calibrating that confidence across more open-ended scientific reasoning remains an unsolved problem.

Looking further ahead, the harder problem may be a tension between two things scientists want at once: hypotheses that are grounded in the literature and free of error, but also genuinely novel. Agents tuned for caution can drift toward the safe and the known; tuned for originality, they are more likely to go astray. This challenge will itself require new ideas, such as better ways to evaluate novelty. Or better ways to connect agents to more specialized models — trained on first-principles scientific data — to help generate ideas that are both original and robust.

Finding the optimal candidate solution

Agents can search for the best solution to a problem, not just a workable one. A materials scientist hunting for a new catalyst faces a near-infinite number of molecular structures, each costly to make and test. A computer scientist looking for a more efficient algorithm faces a similar explosion of possibilities.

Vast solution spaces like these turn up all across science and industry. AlphaEvolve is an agent that can find the best candidates within them. Given a problem expressed as code and a way to score potential solutions, it orchestrates an ensemble of agents to generate many algorithmic candidates, keeps those that score highest, and breeds the next generation from the survivors. It runs unattended, generating and improving candidates at a scale no human team could match.

AlphaEvolve works in code, but its reach extends well beyond software. As Balog points out: "Algorithms can accurately describe so many of the world's scientific and natural processes." AlphaEvolve has assisted in the design of Google's next-generation TPU chips, helped the mathematician Terence Tao solve open Erdős problems, and improved the analysis of genomics data.

But in scientific domains such as materials, AlphaEvolve's top scorers are leads, not final results. A promising catalyst still has to be made and measured at the bench.

Validation

Validation is the slow, costly business of testing whether an idea survives contact with reality. Karl Popper said science advances through conjectures and refutations. Agentic AI is changing the economics of that pairing. AI agents are conjecture machines, making ideas and candidate solutions abundant and relatively cheap. Refutations remain physical and institutional — and so, costly and slow.

> “Algorithms can accurately describe so many of the world's scientific and natural processes.

Mathematics and computer science are often viewed as great exceptions because validation can run *in silico*. An AI agent can generate a proof, represent it in a formal language like Lean, and have the computer verify, unambiguously, that it is correct.

Even for the parts of maths that can't yet be described in formal language, validation is advancing. Aletheia pairs a proof generator with a natural language verifier that checks its work and sends flaws back for revision. In February this year, mathematicians ran the inaugural First Proof challenge: 10 research problems, kept unpublished so they couldn't be found in any training data. In the week allotted, Aletheia solved six — the best result.

Mathematicians will need to absorb all these new outputs. Some already complain that AI-generated proofs are too long and hard to parse (although certain AI-generated proofs are short and elegant). "We are moving toward a future of serious 'proof indigestion' where AI generates breakthroughs faster than humans can review them," says Thang Luong, who led the Aletheia effort. To break this bottleneck, automated verification must become standard, but this verification will need to combine the absolute correctness of formal languages like Lean with the more expressive reasoning of natural language."

Alex Davies, who also leads work on AI for mathematics, is mindful that by automating large chunks of mathematicians' workflows, his discipline is dealing with questions that others may face in the coming years: "I can imagine a world in which machines do the discovery, and what's left for mathematicians is to understand it, and to decide what questions are worth pursuing next." Luong echoes this, noting that maths may also provide some of the general technology needed to address the validation bottleneck in other fields: "One can think of mathematics as an accelerated testbed for the rest of science".

> "We are moving toward a future of serious 'proof indigestion' where AI generates breakthroughs faster than humans can review them.

At the moment, however, the validation gap in most disciplines is widening, not closing. An agent can propose a novel genetic lead to reverse cellular ageing, but cannot say definitively whether it actually works. This explains why companies like Google DeepMind, Ginkgo Bioworks and Lila Sciences are investing in automated labs. But they only suit some fields, are expensive to build and are still early in development. And even automation cannot rush nature's clock. Cell lines need time to grow, chemical reactions take time to complete. For much of science, then, the lab sets the pace.

Implications for policymakers and research funders

To some extent, agents simply add intensity to questions that policymakers are already focused on in their AI for Science strategies, such as how to train the next generation of scientists, how to experiment with new forms of scientific institutions, and how to ensure that AI is not misused by threat actors, while still putting the technology to use addressing the various natural risks that society faces, like the next pandemic.

For every encouraging scenario, there is a challenging one. For example, AI agents could make it more feasible for small, agile teams to pursue creative, ambitious ideas, reversing the trend towards "big team science", or enable scientists to work across domains, bringing new perspectives to existing problems. Assuming efficiency gains make agents sufficiently cost-effective, these trends could particularly benefit smaller, less well-resourced countries and institutions.

But agents will also give rise to anxiety among junior scientists that their institutions are choosing to spend budgets on tokens instead of staff. And left unmanaged, there is a risk that agentic tools could de-skill new generations of scientists before they develop the judgement needed to use them effectively. For the same reason that mathematics students still prove theorems unaided, science-graduate training may need structured periods of agent-free work and access to agents that act as genuine cognitive partners rather than oracles.

Beyond these questions, **AI agents present at least four urgent new priorities: 1.** Scientists need access to the tools. **2.** The tools need access to agent-ready data. **3.** We need more experimental infrastructure to validate AI ideas. **4.** And we need to update the peer review process.

1 Ensure widespread access to agents

Agents will be extremely useful and fallible in non-obvious ways. Both factors provide a strong rationale for policymakers to ensure that all scientists can access the best agents — to speed up discovery and to provide the independent evaluations of AI agents that the scientific community needs to judge how best to use them.

This is an urgent strategic priority for policymakers and science funders, akin to the historical challenge of providing access to supercomputers. Geopolitical debates today often focus on one aspect of sovereign capability — whether a state can train its own frontier model. Much less attention is paid to what may prove to be a more important issue: a country's ability to deploy agents across its scientific ecosystems for transformative impact.

At the micro level, funders must first decide how labs and researchers select and pay for agents. Selection is the easier near-term problem: let researchers find the most useful tools for themselves, without excessive approvals or complex procurement.

Paying is harder. The temptation is to use existing structures, with scientists seeking funding through grant applications or drawing on lab budgets. But the compute required to run agents can be large and the frontier of what agents can do is constantly expanding. While the cost per unit of AI capability is falling fast, total lab expenditures on agents are still likely to rise as agents take on longer and more complex tasks. Policymakers must quickly assess whether budget uplifts or entirely new funding programmes are needed. Delivering access at the scale and price needed will require novel public-private partnerships; the US Genesis Mission is one promising model.

Bigger questions await. Agents may make the existing process of allocating national budgets across disciplines more legible, forcing science funders and research programmes to quantify the investment in data and compute needed to make progress on specific problems. This in turn may lead to more targeted debates about the relative value of solving different problems. If agents propose the top hypotheses to explore across an entire field, with very expensive experimental validation plans, how should this fit into national funding strategies?

2 Make national data assets agent-ready

While scientists need access to agents, agents need access to data. Data that is open or low-risk should be exposed to agents through well-documented APIs, with sufficient quality control and metadata. The engineering support and maintenance to do so is not trivial, so funders should ensure such data stewardship is properly resourced and support interoperable data standards. But ultimately, the ability of agents to help extract and annotate data — from PDFs to download portals — provides an opportunity for governments to extract a lot more value from the data they have already funded.

More sensitive datasets in genomics, virology, or other areas carrying dual-use risk often come with restrictions on who may use them and how. The challenge now is to develop similar privacy-preserving solutions, when appropriate, for agents, with auditability and privacy built in. Examples like OpenSAFELY, which lets human researchers securely access valuable health data, can provide inspiration. The prize is large. A dataset analysis that currently takes years of doctoral work could, with the right secure agent infrastructure, run autonomously in days.

Perhaps most importantly, agents provide a strong rationale for funding the creation of entirely new open datasets. This leads to a further question for funders: could agents help identify the most important datasets to fund? Some of the authors of this article recently made a human-expert-driven attempt to answer that question or fusion energy. How soon will agents be capable of running similar "AI data stocktake" exercises?

3 Tackle the validation bottleneck

Many scientists already struggle to get enough time in facilities to run their experiments. As AI agents make hypotheses and candidate solutions increasingly abundant, this bottleneck will only tighten. Policymakers and funders should address this in at least two ways: investing in existing experimental validation infrastructure and accelerating progress on automated labs.

Public research bodies hold extensive experimental facilities across almost every scientific field. AI agents provide a reinvigorated case for investing in them and opening them up, by renting bench space or experimental run-time to researchers testing computational hypotheses and predictions against reality. Direct partnerships with AI labs are another avenue. Google DeepMind has created a wet lab inside the UK's Francis Crick Institute, a leader in biomedical research, and is also providing independent scientists funding — alongside Co-Scientist access — to carry out the wet lab experiments needed to validate agent-enabled hypotheses. The US government's Genesis Mission will connect the world-class experimental facilities of the Department of Energy's (DOE) National Laboratories with academia and the AI industry.

Automated labs are another promising route to tackling the validation bottleneck, but they currently rely on expensive robotics and compute. Ensuring broad access will require public investment, and there are good early efforts here. The US National Science Foundation has put $100m towards a national network of distributed facilities, while the UK launched a call for ideas and already hosts the £81 million Materials Innovation Factory. The build-out of automated labs will almost certainly go beyond what any single lab or institution can afford, so governments should also explore building centralised capacity and adopting the kind of "user facility" access model seen at the US DOE's national labs.

4 Empower peer reviewers with agents

The peer review process has long been under strain, with slow timelines and reviews of varied quality. Now, scientists are using AI to write ever more grant applications and papers. This is making it harder for funders to know which research to fund, and for peer reviewers to validate findings and identify the most important work.

As noted by Professors James Wilsdon and Geraint Rees, the challenge is not just an increase in supply; writing quality is also no longer a reliable discriminator. Agents deepen the problem. The more an agent is left to plan and optimise an application, drawing on the funder's criteria and its recent winners, the less the bid reflects a scientist's original thinking.

Funders, journals, and conferences are trialling various responses. For grant applications, some suggest leaning less on the written word and more on the investigator's track record and team. The UK's Medical Research Council recently reinstated interviews for shortlisted applicants. These ideas have promise, but also risk privileging seasoned experts or increasing costs.

The likely answer will be a layered approach. Those developing and using AI should document its use more clearly. That could include advancing watermarking techniques and ideas such as "Human-AI Interaction Cards"—short records detailing the prompts and outputs that produced the key scientific insights.

Reviewers should also be able to use agents. This will not be easy, as AI use is often banned, even if widely practised in secret. To move forward, organisations can build on the more nuanced guidelines that some have started to develop and test agents in more objective areas where they are likely to be strong — such as detecting errors.

They could also take steps to ensure that scientists view agents as "human-centric" tools that enhance, rather than bypass, their judgement —f or example, by being transparent about the systems and ensuring that agents expose their reasoning, verify their claims with citations, and (to the extent possible) document their uncertainty. By understanding how an AI agent reaches its conclusion, scientists will have opportunities to learn, intervene, and collaborate.

These are formative years

The scientists using AI agents are not going back. The technology will continue to improve. The institutions that govern science — and the physical and social infrastructure it runs on — were built for a research enterprise that is now evolving faster than they are. We need to upgrade them for the agent era.

This will take a significant collective effort and serious investment. AI agents must not be viewed as a rationale to spend less on science; this would be a tragic false economy. The countries that treat this as a moment of genuine structural change will unlock the greatest public value and shape what comes next. The choices made in this window may be difficult to reverse.

Acknowledgements

We would like to thank the following experts at Google DeepMind who shared insights with us through interviews and feedback on the draft. All mistakes belong to the authors.

Vivek Natarajan, Sebastian Nowozin, Tim Green, Natasha Latysheva, Simon Batzner, Matej Balog, Samuel Albanie, Alex Davies, Nenad Tomasev, Juan Mateos-Garcia, Catherine Pollard, Uchechi Okereke, Agata Laydon, Anna Koivuniemi, Andy Song and Tom Rachman.