



November 2024

A new golden age of discovery

Seizing the *AI for Science* opportunity

Conor Griffin | Don Wallace | Juan Mateos-Garcia | Hanna Schieve | Pushmeet Kohli





2024 年 11 月

一个新的

发现的黄金时代

抓住人工智能促进科学发展的机会

康纳·格里芬 | 唐·华莱士 | 胡安·马特奥斯-加西亚 | 汉娜·希夫 | 普什梅特·科利



Acknowledgements

Thank you to Louisa Bartolo, Zoë Brammer and Nick Swanson for research support, and to the following individuals who shared insights with us through interviews and/or feedback on the draft. All views, and any mistakes, belong solely to the authors.

Žiga Avsec, Nicklas Lundblad, John Jumper, Matt Clifford, Ben Southwood, Craig Donner, Joëlle Barral, Tom Zahavy, Been Kim, Sebastian Nowozin, Matt Clancy, Matej Balog, Jennifer Beroshi, Nitarshan Rajkumar, Brendan Tracey, Yannis Assael, Massimiliano Ciaramita, Michael Webb, Agnieszka Grabska-Barwinska, Alessandro Pau, Tom Lue, Agata Laydon, Anna Koivuniemi, Abhishek Nagaraj, Harry Law, Tom Westgarth, Guy Ward-Jackson, Arianna Manzini, Stefano Bianchini, Sameer Velankar, Ankur Vora, Sébastien Krier, Joel Z Leibo, Elisa Lai H. Wong, Ben Johnson, David Osimo, Andrea Huber, Dipanjan Das, Ekin Dogus Cubuk, Jacklynn Stott, Kelvin Guu, Kiran Vodrahalli, Sanil Jain, Trieu Trinh, Rebeca Santamaria-Fernandez, Remi Lam, Victor Martin, Neel Nanda, Nenad Tomasev, Obum Ekeke, Uchechi Okereke, Francesca Pietra, Rishabh Agarwal, Peter Battaglia, Anil Doshi, Yian Yin, Sukhdeep Singh. ■

致谢

感谢 Louisa Bartolo、Zoë Brammer 和 Nick Swanson 的研究支持，并感谢以下通过采访和/或草案反馈与我们分享见解的人士。所有观点和任何错误仅属于作者。

齐加·阿夫塞克,尼克拉斯·伦德布莱德,约翰·詹珀,马特·克利福德,本·索斯伍德,克雷格·唐纳,乔尔·巴拉尔,汤姆·扎哈维,Been Kim,塞巴斯蒂安·诺沃津,马特·克兰西,马泰·巴洛格,詹妮弗·贝罗什,尼塔山·拉吉库马尔,布伦丹·特雷西,雅尼斯·阿萨尔,马西米利亚诺·西亚拉米塔,迈克尔·韦伯,阿格涅斯卡·格拉布斯卡-巴文斯卡,亚历山德罗·鲍、汤姆·卢、阿加塔·雷登、安娜·科伊维涅米、阿布舍克·纳加拉吉、哈利·劳、汤姆·韦斯特加斯、盖伊·沃德-杰克逊、阿里安娜·曼齐尼、斯特凡诺·比安基尼、萨米尔·韦兰卡、安库尔·沃拉、塞巴斯蒂安·克里尔、乔尔·Z·雷博、Elisa Lai H. Wong, 本·约翰逊, 大卫·奥西莫, 安德里亚·胡贝尔, 迪潘詹·达斯, 艾金·道格斯Cubuk、Jacklynn Stott、Kelvin Guu、Kiran Vodrahalli、Sanil Jain、Trieu Trinh、Rebeca Santamaria-Fernandez、Remi Lam、Victor Martin、Neel Nanda、Nenad Tomasev、Obum Ekeke、Uchechi Okereke、Francesca Pietra、Rishabh Agarwal、Peter Battaglia、Anil多西、伊安·尹、苏赫迪普·辛格。



Introduction

介绍

Introduction

A quiet revolution is brewing in labs around the world, where scientists' use of AI is [growing exponentially](#). [One in three postdocs](#) now use large language models to help carry out literature reviews, coding, and editing. In October, the creators of our [AlphaFold 2](#) system, Demis Hassabis and John Jumper became [Nobel](#) Laureates in Chemistry for using AI to predict the structure of proteins, alongside the scientist David Baker, for his work to design new proteins. Society will soon start to feel these benefits more directly, with [drugs](#) and [materials](#) designed with the help of AI currently making their way through development.

In this essay, we take a tour of how AI is transforming scientific disciplines from genomics to computer science to weather forecasting. Some scientists are training their own AI models, while others are fine-tuning existing AI models, or using these models' predictions to accelerate their research. Scientists are using AI as a scientific instrument to help tackle important problems, such as [designing proteins that bind more tightly to disease targets](#), but are also gradually transforming how science itself is practised.

There is a growing imperative behind scientists' embrace of AI. In recent decades, scientists have continued to deliver consequential advances, from Covid-19 vaccines to renewable energy. But it takes [an ever larger number of researchers to make these breakthroughs](#), and to [transform them into downstream applications](#). As a result, even though the scientific workforce has grown significantly over the past half-century, [rising more than seven fold](#) in the US alone, the societal progress that we would expect to follow, has slowed. For instance, much of the world has witnessed a [sustained slowdown](#) in productivity growth that is undermining the quality of public services. Progress towards the 2030 Sustainable Development Goals, which capture the biggest challenges in health, the environment, and beyond, is [stalling](#).

In particular, scientists looking to make breakthroughs today increasingly run into challenges relating to *scale* and *complexity*, from the ever-growing literature base they need to master, to the increasingly complex experiments they want to run. [Modern deep learning methods](#) are particularly well-suited to these [scale and complexity challenges](#) and can compress the time that future scientific progress would otherwise require. For instance, in structural biology, a single x-ray crystallography experiment to determine the structure of a protein [can take years of work and cost approximately \\$100,000](#), depending on the protein. The [AlphaFold Protein Structure Database](#) now provides instant access to 200 million predicted protein structures for free.

The potential benefits of AI to science are not guaranteed. A significant share of scientists already use LLM-based tools to assist with everyday tasks, such as coding and editing, but the share of scientists using AI-centric research approaches [is much lower, albeit rising rapidly](#). In the rush to use AI, some early scientific use cases have had [questionable impact](#). Policymakers can help accelerate AI's use and steer it towards higher-impact areas. The [US Department of Energy](#), the [European Commission](#), the UK's [Royal Society](#), and the [US National Academies](#), among others, have recently recognised the *AI for Science* opportunity. But no country has yet put a comprehensive strategy in place to enable it.

介绍

世界各地的实验室正在悄然酝酿一场革命，科学家对人工智能的使用呈指数级增长。三分之一的博士后现在使用大型语言模型来帮助进行文献综述、编码和编辑。10月，我们的AlphaFold 2系统的创建者Demis Hassabis和John Jumper因利用人工智能预测蛋白质结构而获得诺贝尔化学奖，与科学家David Baker一起因其设计新蛋白质的工作而获得诺贝尔化学奖。社会很快将开始更直接地感受到这些好处，在人工智能的帮助下设计的药物和材料目前正在开发中。

在本文中，我们将探讨人工智能如何改变科学学科，从基因组学到计算机科学再到天气预报。一些科学家正在训练自己的人工智能模型，而另一些科学家正在微调现有的人工智能模型，或者利用这些模型的预测来加速他们的研究。科学家们正在利用人工智能作为一种科学工具来帮助解决重要问题，例如设计与疾病靶标更紧密结合的蛋白质，但也逐渐改变科学本身的实践方式。

科学家们对人工智能的拥抱越来越迫切。近几十年来，从Covid-19疫苗到可再生能源，科学家们不断取得重大进展。但需要越来越多的研究人员才能取得这些突破，并将其转化为下游应用。因此，尽管科学劳动力在过去半个世纪中显著增长，仅在美国就增长了七倍以上，但我们预期的社会进步却放缓了。例如，世界许多地区的生产率增长持续放缓，从而损害了公共服务的质量。2030年可持续发展目标涵盖了健康、环境等方面的最大挑战，但该目标的进展却陷入停滞。

特别是，如今寻求突破的科学家越来越多地遇到与规模和复杂性相关的挑战，从他们需要掌握的不断增长的文献基础，到他们想要进行的日益复杂的实验。现代深度学习方法特别适合应对这些规模和复杂性挑战，并且可以压缩未来科学进步所需的时间。例如，在结构生物学中，确定蛋白质结构的一次X射线晶体学实验可能需要数年的时间，成本约为100,000美元，具体取决于蛋白质。AlphaFold蛋白质结构数据库现在免费提供对2亿个预测蛋白质结构的即时访问。

人工智能对科学的潜在好处并不能得到保证。很大一部分科学家已经使用基于LLM的工具来协助完成日常任务，例如编码和编辑，但使用以人工智能为中心的研究方法的科学家比例要低得多，尽管在迅速上升。在人工智能的热潮中，一些早期的科学用例产生了值得怀疑的影响。政策制定者可以帮助加速人工智能的使用，并将其引向影响力更大的领域。美国能源部、欧盟委员会、英国皇家学会和美国国家科学院等机构最近都认识到了人工智能促进科学发展的机会。但尚未有国家制定全面战略来实现这一目标。

We hope our essay can inform such a strategy. It is aimed at those who make and influence science policy, and funding decisions. We first identify 5 opportunities where there is a growing imperative to use AI in science and examine the primary ingredients needed to make breakthroughs in these areas. We then explore the most commonly-cited risks from using AI in science, such as to scientific creativity and reliability, and argue that AI can ultimately be net beneficial in each area. We conclude with four public policy ideas to help usher in a new golden age of AI-enabled science.

Throughout the essay we draw on insights from over two dozen interviews with experts from our own AI for Science projects, as well as external experts. The essay naturally reflects our vantage point as a private sector lab, but we believe the case we make is relevant for the whole of science. We hope that readers will respond by sharing their take on the most important AI for Science opportunities, ingredients, risks and policy ideas. ■

Part A: The opportunities	06
Part B: The ingredients	13
Part C: The risks	25
Part D: The policy response	32

我们希望我们的文章能够为这样的策略提供信息。它针对的是那些制定和影响科学政策和资助决策的人。我们首先确定了在科学中使用人工智能日益迫切的 5 个机会，并研究了在这些领域取得突破所需的主要要素。然后，我们探讨了在科学中使用人工智能最常见的风险，例如科学创造力和可靠性，并认为人工智能最终可以在每个领域带来净效益。最后，我们提出了四项公共政策构想，以帮助开创人工智能科学的新黄金时代。

在整篇文章中，我们从对来自我们自己的人工智能科学项目的专家以及外部专家的二十多位采访中汲取了见解。这篇文章自然反映了我们作为私营部门实验室的优势，但我们相信我们提出的案例与整个科学相关。我们希望读者能够分享他们对科学人工智能最重要的机遇、成分、风险和政策想法的看法。

A 部分：机遇	06
B 部分：成分	13
C 部分：风险	25
D 部分：政策反应	32

Part A

The

opportunities

A部分

The

机会

The opportunities

Scientists aim to understand, predict, and influence how the natural and social worlds work, to inspire and satisfy curiosity, and to tackle important problems facing society. [Technologies and methods](#), like the microscope, x-ray diffraction, and statistics, are both *products* of science and *enablers* of it. Over the past century, scientists have increasingly relied on these instruments to carry out their experiments and advance their theories. Computational tools and large-scale data analysis have become particularly important, enabling everything from the discovery of the Higgs boson to the mapping of the human genome. From one view, scientists' growing use of AI is a logical extension of this long-running trend. But it may also signal something much more profound - a discontinuous leap in the limits of what science is capable of.

Rather than listing all areas where it is *possible* to use AI, we highlight five opportunities where we think there is an *imperative* to use it. These opportunities apply across disciplines and address a specific bottleneck, related to scale and complexity, that scientists increasingly face at different points in the scientific process, from generating powerful novel hypotheses to sharing their work with the world. ■

5 opportunities to accelerate science with AI



1. Knowledge

Transform how scientists digest and communicate knowledge



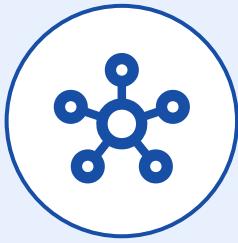
2. Data

Generate, extract, and annotate large scientific datasets



3. Experiments

Simulate, accelerate and inform complex experiments



4. Models

Model complex systems and how their components interact



5. Solutions

Identify novel solutions to problems with large search spaces

机遇

科学家的目标是理解、预测和影响自然世界和社会世界的运作方式，激发和满足好奇心，并解决社会面临的重要问题。显微镜、X射线衍射和统计学等技术和方法既是科学的产物，也是科学的推动者。在过去的一个世纪中，科学家越来越依赖这些仪器来进行实验并推进他们的理论。计算工具和大规模数据分析变得尤为重要，使得从希格斯玻色子的发现到人类基因组图谱的绘制成为可能。从一个角度来看，科学家越来越多地使用人工智能是这一长期趋势的逻辑延伸。但它也可能预示着一些更深刻的事情——科学能力极限的不连续飞跃。

我们没有列出所有可能使用人工智能的领域，而是重点介绍了我们认为有必要使用人工智能的五个机会。这些机会适用于跨学科，并解决与规模和复杂性相关的特定瓶颈，科学家们在科学过程的不同阶段越来越多地面临这一瓶颈，从产生强有力的新颖假设到与世界分享他们的工作。

利用 AI 加速科学发展的 5 个机会 5. 解决方案



1. 知识

改变科学家消化和交流知识的方式



2. 数据

生成、提取和注释大型科学数据集



3. 实验

模拟、加速并提供信息
复杂的实验



4. 模型

对复杂系统及其组件如何交互进行建模



确定问题的新颖解决方案

具有较大的搜索空间

1. Knowledge

Transform how scientists digest and communicate knowledge

To make new discoveries, scientists need to master a pre-existing body of knowledge that continues to grow [exponentially](#) and become ever more specialised. This '[burden of knowledge](#)' helps explain why scientists [making transformative discoveries](#) are increasingly older, interdisciplinary, and located at elite universities, and why the [share of papers](#) authored by individuals, or small teams, is declining, even though [small teams are often better-placed to advance disruptive scientific ideas](#). When it comes to sharing their research there have been welcome innovations such as preprint servers and code repositories, but most scientists still share their findings in [dense, jargon-heavy, English-only papers](#). This can impede rather than ignite interest in scientists' work, including from policymakers, businesses, and the public.

Scientists are already using LLMs, and early scientific assistants based on LLMs, to help address these challenges, such as by [synthesising the most relevant insights](#) from the literature. In an early [demonstration](#), our Science team used our [Gemini](#) LLM to find, extract, and populate specific data from the most relevant subset of 200,000 papers, within a day. Upcoming innovations, such as fine-tuning LLMs on more scientific data and advances in long context windows and citation use, will steadily improve these capabilities. As we expand on below, these opportunities are not without risk. But they provide a window to fundamentally rethink certain scientific tasks, such as what it means to 'read' or 'write' a scientific paper in a world where a scientist can use an LLM to help critique it, tailor its implications for different audiences, or transform it into an 'interactive paper' or [audio guide](#). ■

2. Data

Generate, extract, and annotate large scientific datasets

Despite popular narratives about an era of data abundance, there is a chronic lack of scientific data on most of the natural and social world, from the soil, deep ocean and atmosphere, to the informal economy. AI could help in different ways. It could make existing data collection more accurate, for example by reducing the noise and errors that can occur when [sequencing DNA](#), [detecting cell types in a sample](#), or [capturing animal sounds](#). Scientists can also exploit LLMs' growing ability to operate across images, video and audio, to extract the *unstructured* scientific data that is buried in scientific publications, archives, and less obvious resources such as instructional videos, and convert it into structured datasets.

AI can also help to annotate scientific data with the supporting information that scientists need in order to use it. For example, at least one-third of microbial proteins [are not reliably annotated](#) with details about the function(s) that they are thought to perform. In 2022, our researchers [used AI to predict the function of proteins](#), leading to new entries in the [UniProt](#), [Pfam](#) and [InterPro](#) databases.

1. 知识

改变科学家消化和交流知识的方式

为了做出新的发现，科学家需要掌握现有的知识体系，这些知识体系将继续呈指数级增长并变得更加专业化。这种“知识负担”有助于解释为什么做出变革性发现的科学家年龄越来越大、跨学科且位于精英大学，以及为什么由个人或小团队撰写的论文比例正在下降，尽管小团队往往更好。旨在推进颠覆性的科学思想。当它

在分享他们的研究成果时，出现了一些受欢迎的创新，例如预印本服务器和代码存储库，但大多数科学家仍然以密集、充满行话、纯英文的论文来分享他们的发现。这可能会阻碍而不是激发人们对科学家工作的兴趣，包括政策制定者、企业和公众的兴趣。

科学家们已经在使用 LLMs 和基于 LLMs 的早期科学助手来帮助应对这些挑战，例如通过综合文献中最相关的见解。在早期演示中，我们的科学团队使用 Gemini LLM 在一天之内从 200,000 篇论文的最相关子集中查找、提取和填充特定数据。即将到来的创新，例如对更多科学数据的微调LLMs以及长上下文窗口和引文使用方面的进步，将

稳步提升这些能力。正如我们在下文中详述的那样，这些机会并非没有风险。但它们提供了一个从根本上重新思考某些科学任务的窗口，例如在科学家可以使用 LLM 来帮助批评、定制科学论文的世界中“阅读”或“撰写”科学论文意味着什么。它对不同受众的影响，或将其转化为“交互式纸张”或音频指南。

2. 数据

生成、提取和注释大型科学数据集

尽管流行关于数据丰富时代的说法，但从土壤、深海和大气到非正规经济，大多数自然和社会世界的科学数据长期缺乏。人工智能可以通过不同的方式提供帮助。它可以使现有的数据收集更加准确，例如减少 DNA 测序、检测样本中的细胞类型或捕获动物声音时可能出现的噪音和错误。科学家还可以利用LLMs不断增长的跨图像、视频和音频操作的能力，提取隐藏在科学出版物、档案和教学视频等不太明显的资源中的非结构化科学数据，并将其转换转化为结构化数据集。

人工智能还可以帮助用科学家使用科学数据所需的支持信息来注释科学数据。例如，至少有三分之一的微生物蛋白质没有可靠地注释它们被认为执行的功能的详细信息。2022 年，我们的研究人员使用 AI 来预测蛋白质的功能，从而在 UniProt、Pfam 和 InterPro 数据库中引入新条目。

AI models, once validated, can also serve as new sources of synthetic scientific data. For example, our [AlphaProteo](#) protein design model is trained on more than 100 million AI-generated protein structures from AlphaFold 2, along with experimental structures from the [Protein Data Bank](#). These AI opportunities can complement and increase the return on other much-needed efforts to generate scientific data, such as [digitising archives](#), or funding new data capture technologies and methods, like efforts underway in single cell genomics to create powerful datasets of individual cells in unprecedented detail. ■

3. Experiments

Simulate, accelerate and inform complex experiments

Many scientific experiments are expensive, complex, and slow. Some do not happen at all because researchers cannot access the facilities, participants or inputs that they need. Fusion is a case in point. It promises an energy source that is practically limitless, emission-free and could enable the scaling of energy-intensive innovations, like desalination. To realise fusion, scientists need to create and control *plasma* – a fourth fundamental state of matter. However, the facilities needed are hugely complex to build. [ITER](#)’s prototype tokamak reactor began construction in 2013, but plasma experiments are [not set to begin](#) until the mid-2030s at the earliest, although others hope to build smaller reactors on shorter timelines.

AI could help to simulate fusion experiments and enable much more efficient use of subsequent experiment time. One approach is to run reinforcement learning agents on simulations of physical systems. Between 2019 and 2021, our researchers partnered with the Swiss Federal Institute of Technology Lausanne to [demonstrate](#) how to use RL to control the shape of plasma in a simulation of a tokamak reactor. These approaches could be extended to other experimental facilities, such as [particle accelerators](#), [telescope arrays](#), or [gravitational wave detectors](#).

Using AI to simulate experiments will look very different across disciplines, but a common thread is that the simulations will often inform and guide physical experiments, rather than substitute for them. For example, the average person [has more than 9,000 missense variants](#), or single letter substitutions in their DNA. Most of these genetic variants are benign but some can disrupt the functions that proteins perform, contributing to rare genetic diseases like cystic fibrosis as well as common diseases like cancer. Physical experiments to test the effects of these variants are often limited to a single protein. Our [AlphaMissense model](#) classifies 89% of the 71 million potential human missense variants as likely harmful or benign, enabling scientists to focus their physical experiments on the most likely contributors to disease. ■

人工智能模型一旦经过验证，还可以作为合成科学数据的新来源。例如，我们的 AlphaProteo 蛋白质设计模型经过 AlphaFold 2 中超过 1 亿个 AI 生成的蛋白质结构以及蛋白质数据库中的实验结构的训练。这些人工智能机会可以补充和增加其他急需的科学数据生成工作的回报，例如数字化档案，或资助新的数据捕获技术和方法，例如单细胞基因组学领域正在进行的努力，以前所未有的方式创建强大的单个细胞数据集。细节。

3. 实验

模拟、加速复杂实验并为其提供信息

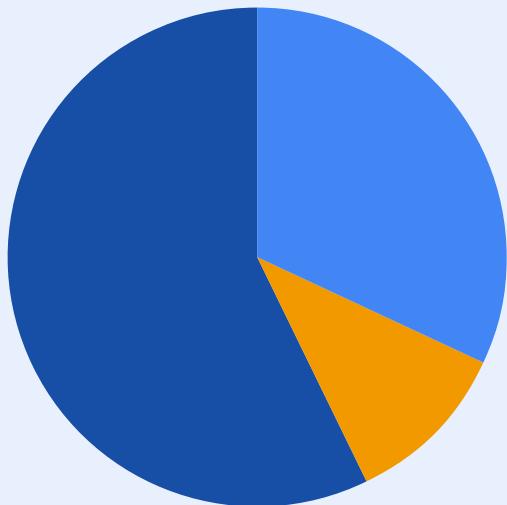
许多科学实验都是昂贵、复杂且缓慢的。有些根本没有发生，因为研究人员无法获得他们所需的设施、参与者或投入。融合就是一个很好的例子。它承诺提供一种几乎无限、零排放的能源，并能够扩大海水淡化等能源密集型创新的规模。为了实现聚变，科学家需要创造和控制等离子体——物质的第四种基本状态。然而，所需的设施建造起来非常复杂。ITER 的原型托卡马克反应堆于 2013 年开始建造，但等离子体实验最早要到 2030 年代中期才会开始，尽管其他人希望在更短的时间内建造更小的反应堆。

人工智能可以帮助模拟聚变实验，并能够更有效地利用后续实验时间。一种方法是在物理系统的模拟上运行强化学习代理。2019 年至 2021 年间，我们的研究人员与瑞士洛桑联邦理工学院合作，演示了如何使用强化学习在托卡马克反应堆模拟中控制等离子体的形状。这些方法可以扩展到其他实验设施，例如粒子加速器、望远镜阵列或引力波探测器。

使用人工智能来模拟实验在不同学科中看起来会非常不同，但一个共同点是模拟通常会为物理实验提供信息和指导，而不是替代它们。例如，普通人的 DNA 中有超过 9,000 个错义变异或单字母替换。这些遗传变异大多数是良性的，但有些会破坏蛋白质的功能，导致囊性纤维化等罕见遗传疾病以及癌症等常见疾病。测试这些变体影响的物理实验通常仅限于单一蛋白质。我们的 AlphaMissense 模型将 7100 万种潜在的人类错义变异中的 89% 分类为可能有害或良性，使科学家能够将物理实验的重点放在最可能导致疾病的因素上。

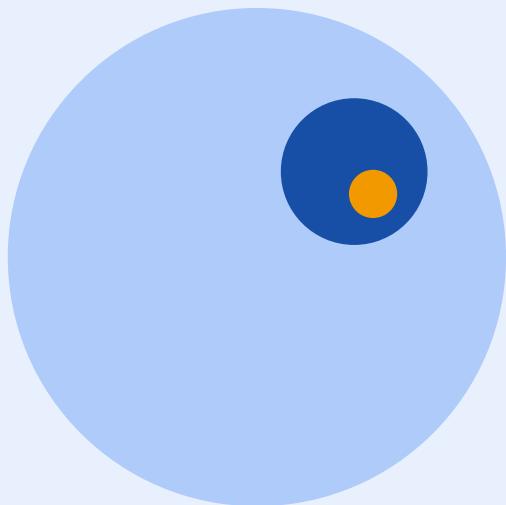
All possible 71 million human missense variants

AlphaMissense predictions:



● Likely benign	57%
● Likely pathogenic	32%
● Uncertain	11%

Human annotations:



● Seen in humans	~6%
● Confirmed by human experts	~0.1%

AlphaMissense predicted the pathogenicity of all possible 71 million missense variants. It classified 89% - predicting 57% were likely benign and 32% were likely pathogenic.

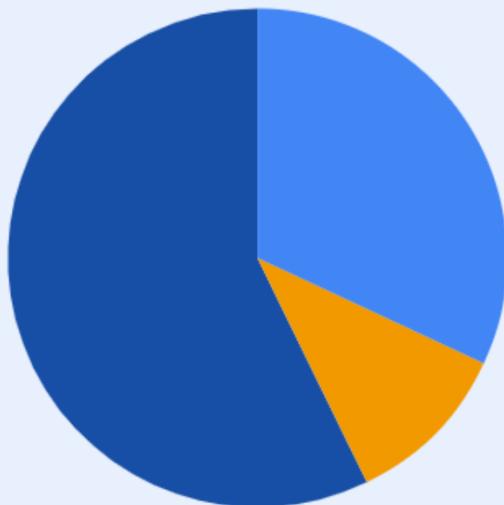
4. Models

Model complex systems and how their components interact

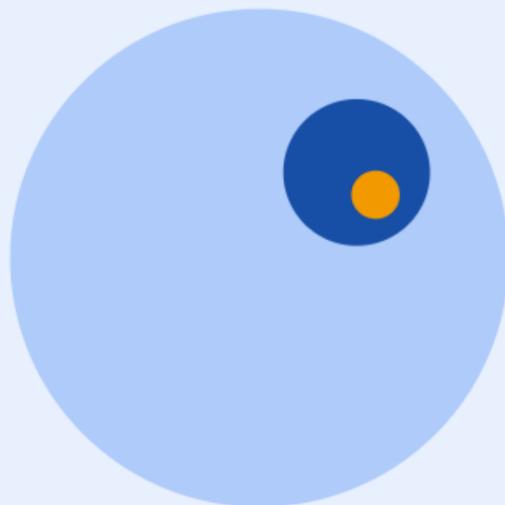
In a [1960 paper](#), the Nobel Prize winning physicist Eugene Wigner marvelled at the “unreasonable effectiveness” of mathematical equations for modelling important natural phenomena, such as planetary motion. However, over the past half century, models that rely on sets of equations or other deterministic assumptions have struggled to capture [the full complexity of systems in biology, economics, weather, and elsewhere](#). This reflects the sheer number of interacting parts that make up these systems, as well as their dynamism and potential for emergent, random or chaotic behaviour. The challenges in modelling these systems impedes scientists’ ability to predict or control how they will behave, including during shocks or interventions, such as rising temperatures, a new drug, or the introduction of a tax change.

所有可能的 7100 万种人类错义变异

AlphaMissense 预测：



人工注释：



AlphaMissense 预测了所有可能的 7100 万个错义变异的致病性。它分类了 89%——预测 57% 可能是良性的，32% 可能是致病的。

4. 型号

对复杂系统及其组件如何交互进行建模

在 1960 年的一篇论文中，诺贝尔奖获得者、物理学家尤金·维格纳 (Eugene Wigner) 对数学方程对于模拟重要自然现象（例如行星运动）的“不合理有效性”感到惊讶。然而，在过去的半个世纪中，依赖方程组或其他确定性假设的模型一直在努力捕捉生物学、经济学、天气和其他领域系统的全部复杂性。这反映了构成这些系统的相互作用部分的绝对数量，以及它们的活力和突发、随机或混乱行为的潜力。对这些系统进行建模的挑战阻碍了科学家预测或控制它们的行为方式的能力，包括在冲击或干预期间，例如气温上升、新药或税收变化的引入。

AI could more accurately model these complex systems by ingesting more data about them, and learning more powerful patterns and regularities within this data. For example, modern weather forecasting is a triumph of science and engineering. For governments and industry, it informs everything from renewable energy planning to preparing for [hurricanes](#) and [floods](#). For the public, the weather is the most popular non-branded query on Google Search. Traditional [numerical prediction methods](#) are based on carefully-defined physics equations that provide a very useful, yet imperfect, approximation of the atmosphere's complex dynamics. They are also computationally expensive to run. In 2023, we released a [deep learning system](#) that predicts weather conditions up to 10 days in advance, which outperformed traditional models on accuracy and prediction speed. As we expand on below, using AI to forecast weather variables could also help to mitigate and respond to climate change. For instance, when pilots fly through humid regions it can cause *condensation trails* that [contribute to](#) aviation's global warming impact. Google scientists [recently used](#) AI to predict when and where humid regions may arise to help pilots avoid flying through them.

In many cases, AI will enrich traditional approaches to modelling complex systems rather than replace them. For example, agent-based modelling simulates interactions between individual actors, like firms and consumers, to understand how these interactions might affect a larger more complex system like the economy. Traditional approaches require scientists to specify beforehand how these computational agents should behave. Our research teams [recently outlined](#) how scientists could use LLMs to create more flexible *generative* agents that communicate and take actions, such as searching for information or making purchases, while also reasoning about and remembering these actions. Scientists could also use reinforcement learning to study how these agents learn and adapt their behaviour in [more dynamic simulations](#), for example in response to the introduction of new energy prices or pandemic response policies. ■

5. Solutions

Identify novel solutions to problems with large search spaces

Many important scientific problems come with a practically incomprehensible number of potential solutions. For example, biologists and chemists aim to determine the structure, characteristics, and function(s) of molecules such as proteins. One goal of such work is to help design novel versions of these molecules to serve as antibody drugs, plastic-degrading enzymes or new materials. However, to design a small molecule drug, scientists face [more than \$10^{60}\$](#) potential options. To design a protein with 400 standard amino acids, they face 20^{400} options. These large search spaces are not limited to molecules but are commonplace for many scientific problems, such as finding the best proof for a maths problem, the most efficient algorithm for a computer science task, or the [best architecture for a computer chip](#).

人工智能可以通过获取更多有关这些复杂系统的数据，并学习这些数据中更强大的模式和规律，来更准确地对这些复杂系统进行建模。例如，现代天气预报是科学和工程的胜利。对于政府和工业界来说，它涵盖了从可再生能源规划到飓风和洪水准备等各个方面。对于公众来说，天气是 Google 搜索上最受欢迎的非品牌查询。传统的数字预测方法基于仔细定义的物理方程，这些方程提供了对大气复杂动力学的非常有用但不完美的近似。它们的运行计算成本也很高。2023 年，我们发布了深度学习系统，可以提前 10 天预测天气状况，在准确性和预测速度上都优于传统模型。正如我们在下文中展开的，使用人工智能预测天气变量也有助于缓解和应对气候变化。例如，当飞行员飞过潮湿地区时，可能会导致凝结尾迹，

加剧航空业对全球变暖的影响。谷歌科学家最近使用人工智能来预测潮湿地区可能出现的时间和地点，以帮助飞行员避免飞过这些地区。

在许多情况下，人工智能将丰富而不是取代传统的复杂系统建模方法。例如，基于代理的建模模拟企业和消费者等个体参与者之间的交互，以了解这些交互可能如何影响更大更复杂的系统（如经济）。传统方法要求科学家事先指定这些计算代理的行为方式。我们的研究团队最近概述了科学家如何使用 LLMs 来创建更灵活的生成代理，这些代理可以进行交流和采取行动，例如搜索信息或进行购买，同时还可以推理和记住这些行动。科学家还可以使用强化学习来研究这些智能体如何在更动态的模拟中学习和调整其行为，例如响应新能源价格或流行病应对政策的引入。



5. 解决方案

确定针对大搜索空间问题的新颖解决方案

许多重要的科学问题都伴随着数量难以理解的潜在解决方案。例如，生物学家和化学家的目标是确定蛋白质等分子的结构、特征和功能。此类工作的目标之一是帮助设计这些分子的新型版本，用作抗体药物、塑料降解酶或新材料。然而，为了设计一种小分子药物，科学家们面临着十多种潜在的选择。为了设计一种含有 400 个标准氨基酸的蛋白质，他们面临 20 个选择。这些大型搜索空间不仅限于分子，而且对于许多科学问题来说都很常见，例如寻找数学问题的最佳证明、计算机科学任务的最有效算法或计算机芯片的最佳架构。

Traditionally, scientists rely on some combination of intuition, trial and error, iteration, or brute force computing to find the best molecule, proof, or algorithm. However, these methods struggle to exploit the huge space of potential solutions, leaving better ones undiscovered. AI can [open up new parts of these search spaces](#) while also homing in more quickly on the solutions that are most likely to be viable and useful - a delicate balancing act. For example, in July, our [AlphaProof and AlphaGeometry 2](#) systems correctly solved four out of six problems from the [International Mathematical Olympiad](#), an elite high school competition. The systems make use of our Gemini LLM architecture to generate a large number of novel ideas and potential solutions for a given maths problem, and combine this with systems grounded in mathematical logic that can iteratively work towards the candidate solutions that are most likely to be correct. ■

AI scientists or AI-empowered scientists?

This growing use of AI in science, and the emergence of early AI scientific assistants, raises questions about how fast and how far the capabilities of AI may advance and what this will mean for human scientists. Current LLM-based AI scientific assistants make a relatively small contribution to a relatively narrow range of tasks, such as supporting literature reviews. There are plausible near-term scenarios in which they become better at these tasks and become capable of more impactful ones, such as helping to generate powerful hypotheses, or helping to predict the outcomes of experiments. However, current systems still struggle with the deeper creativity and [reasoning](#) that human scientists rely on for such tasks. [Efforts are underway to improve these AI capabilities](#), for example by combining LLMs with logical deduction engines, as in our [AlphaProof and AlphaGeometry 2](#) examples, but further breakthroughs are needed. The ability to accelerate or automate experiments will also be harder for those that require complicated actions in wet labs, interacting with human participants, or lengthy processes, such as monitoring disease progression. Although again, work is underway in some of these areas, such as new types of laboratory robotics and [automated labs](#).

Even as AI systems' capabilities improve, the greatest marginal benefit will come from deploying them in use cases that play to their relative strengths - such as the ability to rapidly extract information from huge datasets - and which help address genuine bottlenecks to scientific progress such as the five opportunities outlined above, rather than automating tasks that human scientists already do well. As AI enables cheaper and more powerful science, demand for science and scientists will also grow. For example, recent breakthroughs have already led to a slew of new startups in areas like [protein design](#), [material science](#) and [weather forecasting](#). Unlike other sectors, and despite [past claims to the contrary](#), future demand for science appears practically limitless. New advances [have always opened up new, unpredictable regions](#) in the scientific *map of knowledge*, and AI will do similar. As [envisioned](#) by Herbert Simon, AI systems will also become *objects* of science research, with scientists set to play a leading role in evaluating and explaining their scientific capabilities, as well as in developing new types of human-AI scientific systems. ■

传统上，科学家依靠直觉、反复试验、迭代或强力计算的某种组合来找到最好的分子、证据或算法。然而，这些方法很难利用潜在解决方案的巨大空间，而没有发现更好的解决方案。人工智能可以开辟这些搜索空间的新部分，同时也可以更快地找到最有可能可行和有用的解决方案——这是一种微妙的平衡行为。例如，7月份，我们的 AlphaProof 和 A IphaGeometry 2 系统正确解决了精英高中竞赛国际数学奥林匹克竞赛中的六道题中的四道。该系统利用我们的 Gemini LLM 架构为给定的数学问题生成大量新颖的想法和潜在的解决方案，并将其与基于数学逻辑的系统相结合，这些系统可以迭代地寻找候选解决方案最有可能是正确的。

人工智能科学家还是人工智能科学家？

人工智能在科学中的日益广泛使用，以及早期人工智能科学助理的出现，引发了人们的疑问：人工智能的能力可能会以多快的速度和多远的速度进步，以及这对人类科学家意味着什么。当前基于LLM的人工智能科学助手对范围相对狭窄的任务做出的贡献相对较小，例如支持文献综述。在短期内，他们可能会更好地完成这些任务，并能够完成更有影响力的任务，例如帮助产生强有力假设，或帮助预测实验结果。然而，当前的系统仍然难以应对人类科学家完成此类任务所依赖的更深层次的创造力和推理能力。我们正在努力提高这些人工智能能力，例如将 LLMs 与逻辑推演引擎相结合，如我们的 AlphaProof 和 AlphaGeometry 2 示例中所示，但还需要进一步的突破。对于那些需要在潮湿实验室中进行复杂操作、与人类参与者互动或长时间进行实验的人来说，加速或自动化实验的能力也将变得更加困难。

过程，例如监测疾病进展。尽管如此，其中一些领域的工作正在进行中，例如新型实验室机器人和自动化实验室。

即使人工智能系统的能力不断提高，最大的边际效益仍将来自于将它们部署在发挥其相对优势的用例中——例如从庞大数据库集中快速提取信息的能力——并有助于解决科学进步的真正瓶颈，例如上面概述的五个机会，而不是自动化人类科学家已经擅长的任务。

随着人工智能使科学变得更便宜、更强大，对科学和科学家的需求也将增长。例如，最近的突破已经在蛋白质设计、材料科学和天气预报等领域催生了大量新的初创公司。与其他部门不同的是，尽管过去的说法与此相反，但未来对科学的需求似乎实际上是无限的。新的进步总是在科学知识地图中开辟新的、不可预测的领域，人工智能也会做类似的事情。正如赫伯特·西蒙的设想，人工智能系统也将成为科学的研究对象，科学家将在评估和解释其科学能力以及开发新型人类人工智能科学系统方面发挥主导作用。

Part B

The

ingredients

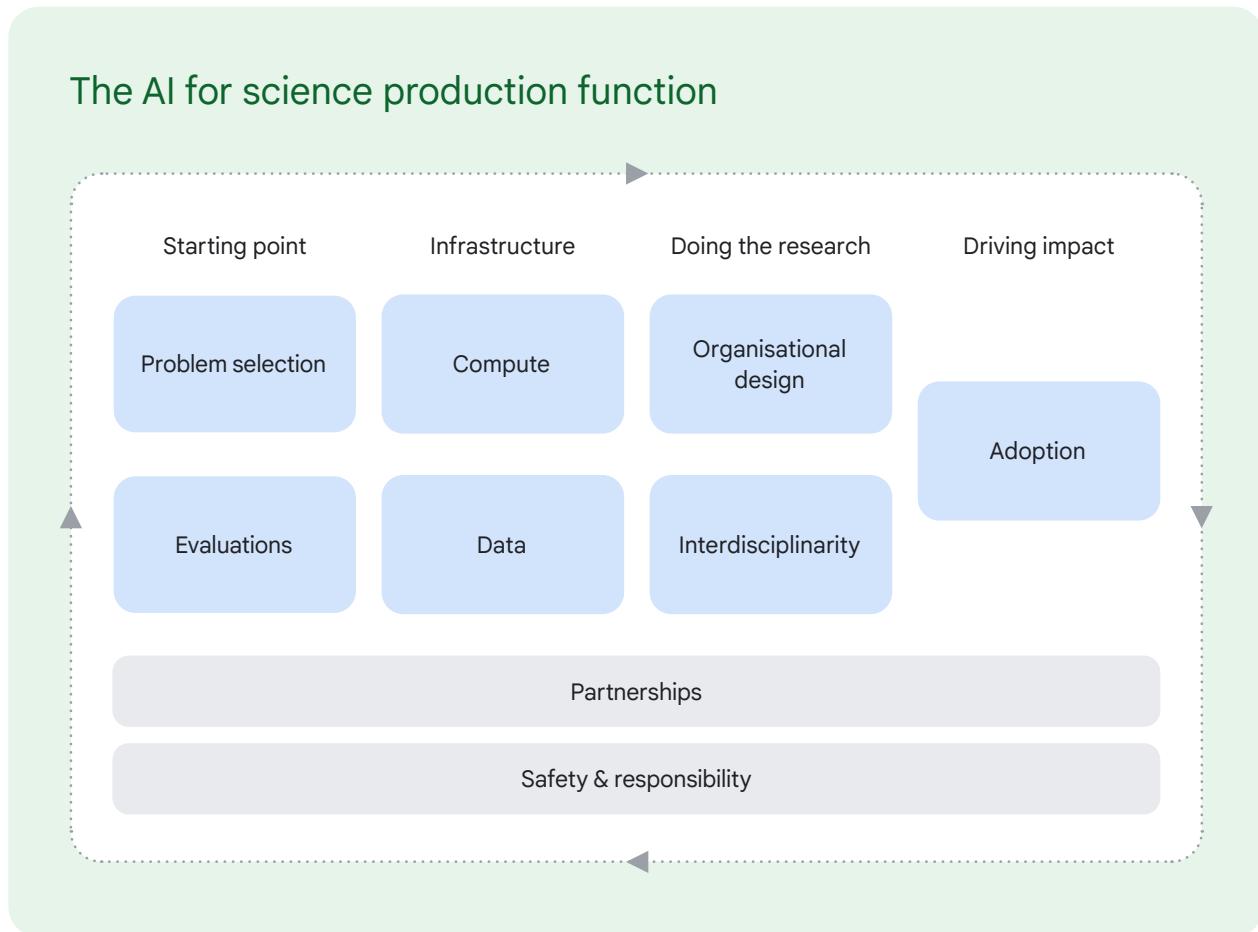
B部分

The

原料

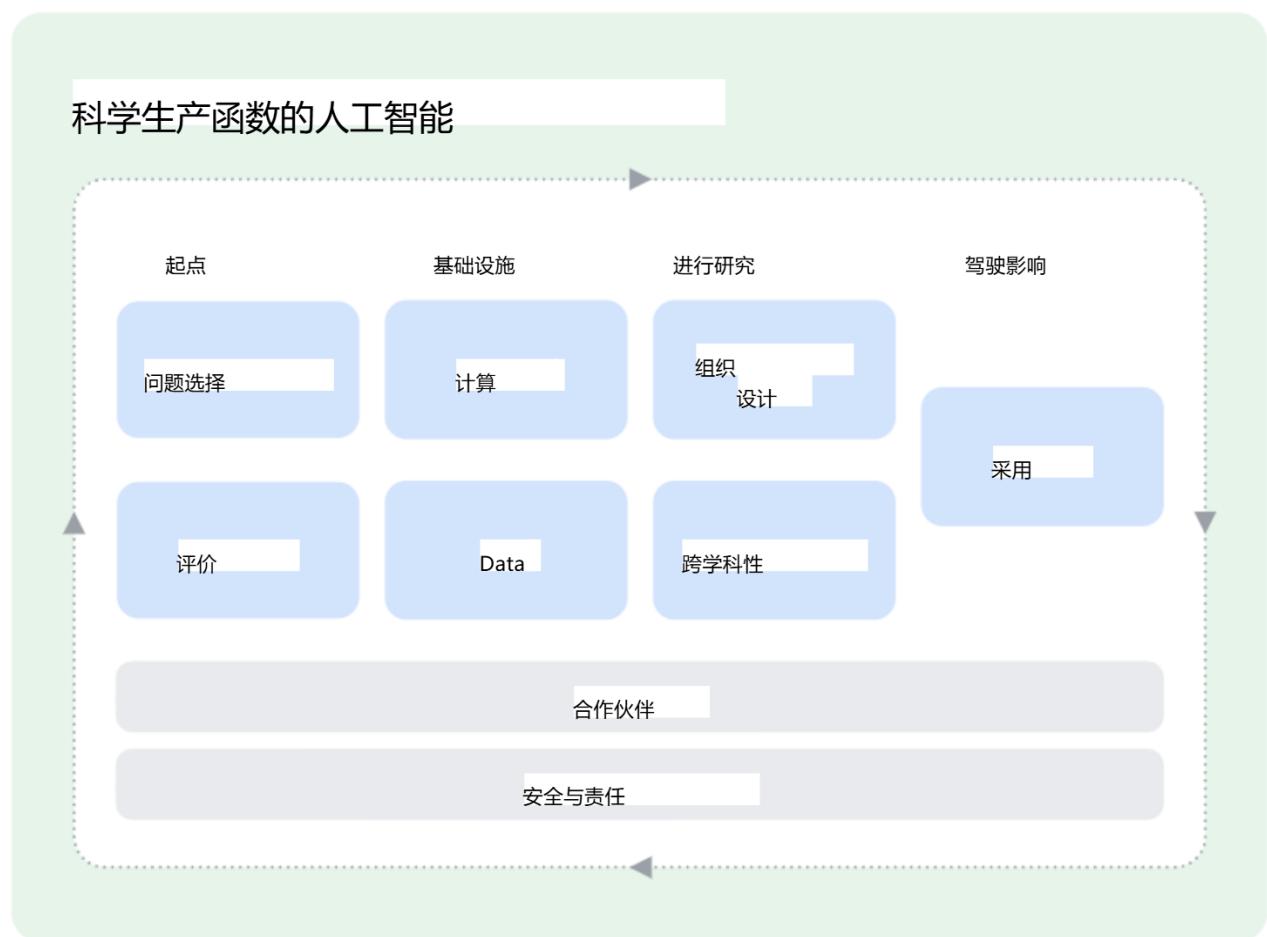
The ingredients

We are interested in the ingredients that ambitious AI for Science efforts need to succeed - both at the individual research effort level and at the level of the science ecosystem, where policymakers have more scope to shape them. The experts that we interviewed routinely cited several ingredients that we organised into a toy model, which we call *the AI for Science production function*. This production function is not meant to be exhaustive, prescriptive, or a neat linear process. The ingredients will be intuitive to many, but our interviews revealed a number of lessons about what they look like in practice which we share below. ■



成分

我们对雄心勃勃的人工智能科学努力取得成功所需的要素感兴趣——无论是在个人研究工作层面还是在科学生态系统层面，政策制定者有更大的空间来塑造它们。我们采访的专家经常引用我们组织成玩具模型的几种成分，我们将其称为科学人工智能生产函数。这个生产函数并不意味着是详尽的、规定性的或简洁的线性过程。这些成分对许多人来说都是直观的，但我们的采访揭示了一些关于它们在实践中的样子的经验教训，我们在下面分享。



1. Problem selection

Pursue ambitious, AI-shaped problems

Scientific progress rests on being able to identify an important problem and ask the *right* question about how to solve it. In [their exploration](#) into the genesis of scientific breakthroughs, Venkatesh Narayananamurti and Jeffrey Y. Tsao document how important the reciprocal and recursive relationship between questions and answers is, including the importance of asking ambitious new questions. Our Science team starts by thinking about whether a potential research problem is significant enough to justify a substantial investment of time and resources. Our CEO Demis Hassabis has a mental model to guide this assessment: thinking about all of science as a *tree of knowledge*. We are particularly interested in the roots - fundamental ‘root node problems’ like [protein structure prediction](#) or [quantum chemistry](#) that, if solved, could unlock entirely new branches of research and applications.

To assess whether AI will be suitable and additive, we look for problems with certain characteristics, such as huge combinatorial search spaces, large amounts of data, and a clear objective function to benchmark performance against. Often a problem is suitable for AI in principle, but the inputs aren’t yet in place and it needs to be stored for later. One of the original [inspirations](#) for AlphaFold was conversations that Demis had many years prior as a student with a friend who was obsessed with the protein folding problem. Many recent breakthroughs also feature this coming together of an important scientific problem and an AI approach that has reached a point of maturity. For example, our [fusion](#) effort was aided by a novel [reinforcement learning algorithm](#) called *maximum a posteriori policy optimization*, which had only just been released. Alongside a new fast and [accurate simulator](#) that our partners EPFL had just developed, that enabled the team to overcome a data paucity challenge.

In addition to picking the right problem, it is important to *specify* it at the right level of difficulty. Our interviewees emphasised that a powerful problem statement for AI is often one that lends itself to *intermediate* results. If you pick a problem that’s too hard then you won’t generate enough signal to make progress. Getting this right relies on intuition and experimentation. ■

1. 问题选择

追求雄心勃勃的人工智能型问题

科学进步取决于能够识别重要问题并提出有关如何解决它的正确问题。在探索科学突破的起源时，Venkatesh Narayananamurti 和 Jeffrey Y. Tsao 记录了问题和答案之间的相互和递归关系的重要性，包括提出雄心勃勃的新问题的重要性。我们的科学团队首先考虑潜在的研究问题是否足够重要，值得投入大量时间和资源。我们的首席执行官 Demis Hassabis 有一个心理模型来指导这一评估：将所有科学视为一棵知识树。我们对根源特别感兴趣，即基本的“根节点问题”，例如蛋白质结构预测或量子化学，如果解决这些问题，可以解锁全新的研究和应用分支。

为了评估人工智能是否适用和可附加，我们寻找具有某些特征的问题，例如巨大的组合搜索空间、大量数据以及用于衡量性能的明确目标函数。通常，一个问题原则上适合人工智能，但输入尚未到位，需要存储起来供以后使用。AlphaFold 的最初灵感之一是 Demis 多年前在学生时代与一位痴迷于蛋白质折叠问题的朋友的对话。最近的许多突破也体现了重要的科学问题和已经达到成熟点的人工智能方法的结合。例如，我们的融合工作得到了一种名为最大后验策略优化的新型强化学习算法的帮助，该算法刚刚发布。除了我们的合作伙伴 EPFL 刚刚开发的新型快速准确模拟器之外，该模拟器还使团队能够克服数据匮乏的挑战。

除了选择正确的问题之外，以正确的难度级别指定问题也很重要。我们的受访者强调，人工智能的有力问题陈述往往有助于产生中间结果。如果你选择的问题太难，那么你将无法产生足够的信号来取得进展。做到这一点依赖于直觉和实验。

2. Evaluations

Invest in evaluation methods that can provide a robust performance signal and are endorsed by the community

Scientists use evaluation methods, such as benchmarks, metrics and competitions, to assess the scientific capabilities of an AI model. Done well, these evaluations provide a way to track progress, encourage innovation in methods, and galvanise researchers' interest in a scientific problem. Often, a variety of evaluation methods are required. For example, our [weather forecasting](#) team started with an initial 'progress metric' based on a few key variables, such as surface temperature, that they used to 'hill climb', or gradually improve their model's performance. When the model had reached a certain level of performance, they carried out a more comprehensive evaluation using more than 1,300 metrics inspired by the [European Centre for Medium-Range Weather Forecasts's evaluation scorecard](#). In past work, the team learned that AI models can sometimes achieve good scores on these metrics in undesirable ways. For example, 'blurry' predictions - such as predicting rainfall within a large geographical area - are less penalised than 'sharp' predictions - such as predicting a storm in a location that is very slightly different to the actual location - the so-called 'double-penalty' problem. To provide further verification, the team evaluated the usefulness of their model on downstream tasks, such as its ability to predict the track of a cyclone, and to characterise the strength of 'atmospheric rivers' - narrow bands of concentrated moisture that can lead to flooding.

The most impactful AI for Science evaluation methods are often community-driven or endorsed. A gold standard is the [Critical Assessment of protein Structure Prediction](#) competition. Established in 1994 by [Professor John Moult and Professor Krzysztof Fidelis](#), the biennial CASP competition has challenged research groups to test the accuracy of their protein structure prediction models against real, unreleased, experimental protein structures. It has also become a unique global community and a catalyst for research progress, albeit one that is hard to replicate quickly. The need for community buy-in also provides an argument for publishing benchmarks so that researchers can use, critique and improve them. However, this also creates the risk that the benchmark will 'leak' into an AI model's training data, reducing its usefulness for tracking progress. There is no perfect solution to this tradeoff but, at a minimum, new public benchmarks are needed at regular intervals. Scientists, AI labs and policymakers should also explore new ways to assess the scientific capabilities of AI models, such as setting up new third-party assessor organisations, competitions, and enabling more open-ended probing of AI models' capabilities by scientists. ■

2. 评价

投资能够提供稳健绩效信号并得到社区认可的评估方法

科学家使用基准、指标和竞赛等评估方法来评估人工智能模型的科学能力。如果做得好，这些评估可以提供一种跟踪进展、鼓励方法创新并激发研究人员对科学问题的兴趣的方法。通常，需要多种评估方法。例如，我们的天气预报团队从基于一些关键变量（例如表面温度）的初始“进度指标”开始，他们用这些变量“爬山”或逐渐提高模型的性能。当模型达到一定的性能水平时，他们受到欧洲中期天气预报中心评估记分卡的启发，使用 1,300 多个指标进行了更全面的评估。在过去的工作中，团队了解到人工智能模型有时会以不受欢迎的方式在这些指标上取得良好的分数。例如，“模糊”预测（例如预测大范围地理区域内的降雨量）比“清晰”预测（例如预测与实际位置略有不同的位置发生风暴）受到的惩罚要小，即所谓的“精确”预测。“双重惩罚”问题。为了提供进一步的验证，该团队评估了模型在下游任务中的有用性，例如其预测气旋路径的能力，以及表征“大气河流”强度的能力——可能导致洪水的集中水分窄带。

最有影响力的科学人工智能评估方法通常是社区驱动或认可的。黄金标准是蛋白质结构预测竞赛的批判性评估。CASP 竞赛由 John Moult 教授和 Krzysztof Fidelis 教授于 1994 年创办，两年一度的 CASP 竞赛要求研究小组根据真实的、未发布的实验蛋白质结构来测试其蛋白质结构预测模型的准确性。它也已成为一个独特的全球社区和研究进步的催化剂，尽管很难快速复制。社区支持的需要也为发布基准提供了理由，以便研究人员可以使用、批评和改进它们。然而，这也带来了基准测试“泄漏”到人工智能模型的训练数据中的风险，从而降低了其跟踪进度的有用性。对此没有完美的解决方案

但至少需要定期制定新的公共基准。科学家、人工智能实验室和政策制定者还应该探索评估人工智能模型科学能力的新方法，例如建立新的第三方评估组织、竞赛，以及让科学家对人工智能模型的能力进行更开放的探索。



3. Compute

Track how compute use is evolving and invest in specialist skills

Multiple [government reviews have recognised](#) the growing importance of compute to progress in AI, science, and the wider economy. As we expand on further below, there is also a growing focus on its energy consumption and greenhouse gas emissions. AI labs and policymakers should take a grounded, long-term view that considers how compute needs will vary across AI models and use cases, potential multiplier effects and efficiency gains, and how this compares to counterfactual approaches to scientific progress that don't use AI.

For example, some state-of-the-art AI models, such as in protein design, are relatively small. Larger models, like LLMs, are compute-intensive to train but typically require much less compute to fine-tune, or to run inference against, which can open up more efficient pathways to science research. Once an LLM is trained, it is also easier to make it more efficient, for example via better [data curation](#), or by '[distilling](#)' the large model into a smaller one. AI compute needs should also be evaluated in comparison to other models of scientific progress. For example, AI weather forecasting models are compute-intensive to train, but [can still be](#) more computationally-efficient than traditional techniques. These nuances highlight the need for AI labs and policymakers to track compute use empirically, to understand how it is evolving, and to project what these trends mean for future demand. In addition to ensuring sufficient access to the right kind of chips, a compute strategy should also prioritise the critical infrastructure and engineering skills needed to manage access and ensure reliability. This is often under-resourced in academia and public research institutions. ■

4. Data

Blend top-down and bottom-up efforts to collect, curate, store, and access data

Similar to compute, data can be viewed as critical *infrastructure* for AI for Science efforts that needs to be developed, maintained, and updated over time. Discussions often focus on identifying new datasets that policymakers and practitioners should create. There is a role for such top-down efforts. In 2012, the Obama Administration launched the [Materials Project](#) to map known and predicted materials, such as inorganic crystals, like silicon, that are found in batteries, solar panels, and computer chips. Our recent [GNoME](#) effort used this data to predict 2.2 million novel inorganic crystals, including 380,000 that simulations suggest are stable at low temperatures, making them candidates for new materials.

However, it is often difficult to predict in advance what scientific datasets will be most important, and many AI for Science breakthroughs rely on data that emerged more organically, thanks to the efforts of an enterprising individual or small teams. For example, [Daniel MacArthur](#), then a researcher at the Broad Institute, led the development of the [gnomAD](#) dataset of genetic variants that our [AlphaMissense](#) work subsequently drew on. Similarly, the mathematical proof assistant and programming language [Lean](#) was originally developed by the programmer [Leonardo de Moura](#). It is not a dataset, but many AI labs now use it to help train their AI maths models, including our [AlphaProof](#) system.

3. 计算

跟踪计算使用的演变并投资于专业技能

多项政府审查已经认识到计算对于人工智能、科学和更广泛的经济进步日益重要。随着我们在下面进一步展开，人们也越来越关注其能源消耗和温室气体排放。人工智能实验室和政策制定者应采取脚踏实地的长期观点，考虑计算需求在人工智能模型和用例之间的变化、潜在的乘数效应和效率增益，以及与不使用人工智能的反事实科学进步方法相比如何。

例如，一些最先进的人工智能模型（例如蛋白质设计）相对较小。较大的模型（例如LLMs）的训练需要大量计算，但通常需要更少的计算来进行微调或运行推理，这可以为科学研究开辟更有效的途径。一旦训练了LLM，就可以更轻松地提高其效率，例如通过更好的数据管理，或将大型模型“提炼”为较小的模型。还应评估人工智能计算需求

与其他科学进步模型的比较。例如，人工智能天气预报模型的训练需要大量计算，但仍然比传统技术具有更高的计算效率。这些细微差别凸显了人工智能实验室和政策制定者需要根据经验跟踪计算使用情况，了解其发展情况，并预测这些趋势对未来需求的影响。除了确保对正确类型芯片的充分访问之外，计算策略还应优先考虑管理访问和确保可靠性所需的关键基础设施和工程技能。学术界和公共研究机构的资源往往不足。



4. 数据

混合自上而下和自下而上的努力来收集、管理、存储和访问数据

与计算类似，数据可以被视为人工智能科学工作的关键基础设施，需要随着时间的推移进行开发、维护和更新。讨论通常集中在确定政策制定者和从业者应该创建的新数据集。这种自上而下的努力是有作用的。2012年，奥巴马政府启动了材料项目，绘制已知和预测材料的地图，例如电池、太阳能电池板和计算机芯片中发现的硅等无机晶体。我们最近的GNoME工作利用这些数据预测了220万种新型无机晶体，其中模拟表明在低温下稳定的380,000种晶体使它们成为新材料的候选者。

然而，通常很难提前预测哪些科学数据集将是最重要的，并且许多人工智能科学突破都依赖于更有机地出现的数据，这要归功于有进取心的个人或小团队的努力。例如，时任博德研究所研究员的丹尼尔·麦克阿瑟(Daniel MacArthur)领导了遗传变异gnomAD数据集的开发，我们的AlphaMissense工作随后借鉴了该数据集。同样，数学证明助手和编程语言Lean最初是由程序员Leonardo de Moura开发的。它不是一个数据集，但许多人工智能实验室现在使用它来帮助训练他们的人工智能数学模型，包括我们的AlphaProof系统。

Efforts like [gnomAD](#) or [Lean](#) highlight how top-down data efforts need to be complemented by better incentives for individuals at all stages of the data pipeline. For example, some data from strategic wet lab experiments is currently discarded, but could be collected and stored, if stable funding was available. Data curation could also be better incentivised. Our [AlphaFold](#) models were trained on data from the [Protein Data Bank](#) that was particularly high quality because journals require the deposition of protein structures as a precondition for publication, and the PDB's professional data curators developed standards for this data. In genomics, many researchers are also obliged to deposit raw sequencing data in the [Sequence Read Archive](#) but inconsistent standards mean that individual datasets often still need to be reprocessed and combined. Some other high-quality datasets go unused altogether, because of restrictive licensing conditions, such as in biodiversity, or because the datasets are not released, such as decades of data from publicly-funded fusion experiments. There can be logical reasons for this, such as a lack of time, funds, somewhere to put the data, or the need for temporary embargo periods for researchers who develop the data. But in aggregate these data access issues pose a key bottleneck to using AI to advance scientific progress. ■

5. Organisational design

Strike the right balance between bottom-up creativity and top-down coordination

A simple heuristic is that academia and industry tend to approach science research at two ends of a spectrum. Academia tends to be more bottom-up, and industry labs tend to be more top-down. In reality, there has long been plenty of space in between, particularly at the most successful labs, such as the golden eras of Bell Labs and Xerox PARC that were renowned for their blue skies research and served as inspiration in the founding of DeepMind. Recently, a new wave of [science research institutions](#) has emerged that try to [learn from](#) these outlier examples. These organisations differ in their goals, funding models, disciplinary focus, and how they organise their work. But collectively they want to deliver more high-risk, high-reward research, less bureaucracy, and better incentives for scientists. Many have a strong focus on applying AI, such as the UK's [Advanced Research and Invention Agency](#), the [Arc Institute](#), and the growing number of [Focused Research Organisations](#) that aim to tackle specific problems in science that are too large for academia and not profitable enough for industry, such as the [organisation](#) tasked with expanding the Lean proof assistant that is pivotal to AI maths research.

At their core, these new institutions share a desire to find a better blend of top-down coordination and focus with bottom-up empowerment of scientists. For some organisations, this means focussing on a single specific problem with pre-specified milestones. For others, it means offering more unrestricted funding to principal investigators. Getting this balance right is critical to attracting and retaining research leaders, who must also buy into it if it is to succeed - Demis Hassabis has credited it as the single biggest factor for successfully coordinating cutting-edge research at scale. Striking this balance is also important *within* individual research efforts. In Google DeepMind's case, efforts often pivot between more unstructured 'exploration' phases, where teams search for new ideas, and faster 'exploitation' phases, where they focus on engineering and scaling performance. There is an art to knowing when to switch between these modes and how to adapt the project team accordingly. ■

像gnomAD或Lean这样的努力凸显了自上而下的数据工作需要如何通过对数据管道各个阶段的个人提供更好的激励来补充。例如，战略湿实验室实验的一些数据目前被丢弃，但如果资金稳定，可以收集和存储。数据管理也可以得到更好的激励。我们的AlphaFold模型是根据蛋白质数据库中的数据进行训练的，这些数据质量特别高，因为期刊需要沉积蛋白质结构作为发表的先决条件，并且PDB的专业数据管理员为这些数据制定了标准。在基因组学中，许多研究人员也有义务将原始测序数据存放在序列读取存档中，但不一致的标准意味着单个数据集通常仍需要重新处理和组合。其他一些高质量数据集由于限制性许可条件（例如生物多样性）而完全未使用，或者因为数据集未发布（例如来自公共资助的融合实验的数十年数据）。

造成这种情况可能有逻辑原因，例如缺乏时间、资金、放置数据的地方，或者开发数据的研究人员需要临时禁运期。但总的来说，这些数据访问问题构成了利用人工智能推动科学进步的关键瓶颈。

5. 组织设计

在自下而上的创造力和自上而下的协调之间取得适当的平衡

一个简单的启发是，学术界和工业界倾向于在光谱的两端进行科学研究。学术界往往更加自下而上，而工业实验室往往更加自上而下。事实上，两者之间长期以来存在着很大的空间，特别是在最成功的实验室，例如贝尔实验室和施乐帕洛阿尔托研究中心的黄金时代，它们以其蓝天研究而闻名，并为DeepMind的创立提供了灵感。最近，出现了新一波科学研究所机构，试图从这些异常例子中学习。这些组织的目标不同，

资助模式、学科重点以及他们如何组织工作。但总的来说，他们希望为科学家提供更多高风险、高回报的研究，减少官僚主义，并提供更好的激励。许多组织都非常关注人工智能的应用，例如英国高级研究和发明局、Arc研究所，以及越来越多的重点研究组织，这些组织旨在解决科学领域的具体问题，这些问题对学术界来说太大，而且利润不足以使学术界带来足够的利润。行业，例如负责扩展对人工智能数学研究至关重要的精益证明助手的组织。

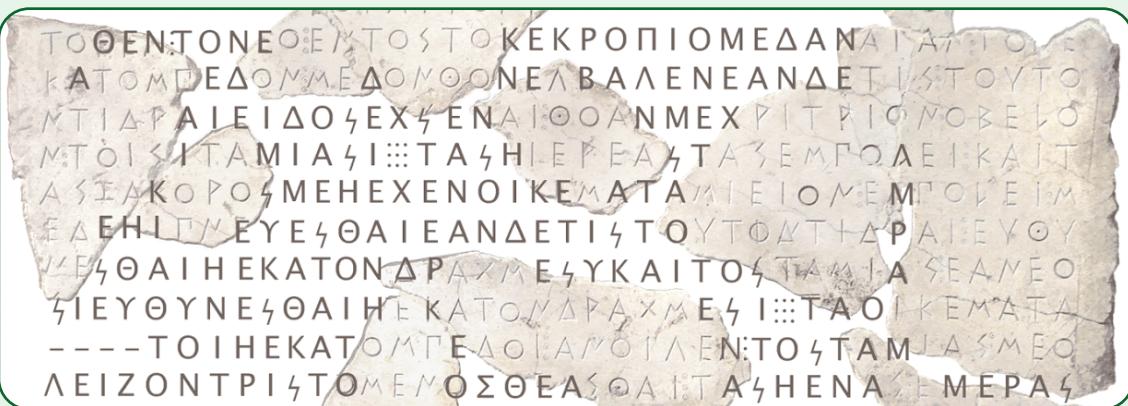
从本质上讲，这些新机构的共同愿望是找到自上而下的协调和重点与自下而上的科学家赋权的更好结合。对于某些组织来说，这意味着通过预先指定的里程碑专注于单个特定问题。对于其他人来说，这意味着向主要研究人员提供更多不受限制的资金。实现这种平衡对于吸引和留住研究领导者至关重要，如果要取得成功，研究领导者也必须接受它——杰米斯·哈萨比斯将其视为成功协调大规模尖端研究的最大因素。在个人研究工作中实现这种平衡也很重要。在Google DeepMind的案例中，工作通常在更加非结构化的“探索”阶段和更快的“开发”阶段之间进行，在“探索”阶段，团队寻找新想法，在更快的“开发”阶段，他们专注于工程和扩展性能。知道何时在这些模式之间切换以及如何相应地调整项目团队是一门艺术。

6. Interdisciplinarity

Approach science as a team, fund neglected roles, and promote a culture of contestability

Many of the hardest scientific problems require progress at the boundaries between fields. However when practitioners are brought together, [for example during Covid-19](#), they often struggle to transition from *multidisciplinary* teams - where they each retain their own disciplinary angle - to genuine *interdisciplinarity*, where they collectively develop [shared ideas and methods](#). This challenge reflects the growing specialisation of scientific knowledge, as well as [incentives](#) such as grant funding, that often evaluate practitioners predominantly on their core expertise.

AI for Science efforts are often multidisciplinary by default, but to succeed, they need to become genuinely interdisciplinary. A starting point is to pick a problem that requires each type of expertise, and then provide enough time and focus to cultivate a team dynamic around it. For example, our [Ithaca](#) project used AI to restore and attribute damaged ancient Greek inscriptions, which could help practitioners to study the thought, language, and history of past civilizations. To succeed, project co-lead Yannis Assael had to [develop](#) an understanding of *epigraphy* - the study of ancient inscribed text. The project's epigraphers, in turn, had to learn how the AI model worked, given the importance of intuition to their work. Cultivating these team dynamics requires the right incentives. Empowering a small, tight-knit team to focus on solving *the problem*, rather than authorship of papers, was key to the AlphaFold 2 breakthrough. This type of focus can be easier to achieve in industry labs, but again highlights the importance of longer-term public research funding that is less tied to publication pressures.



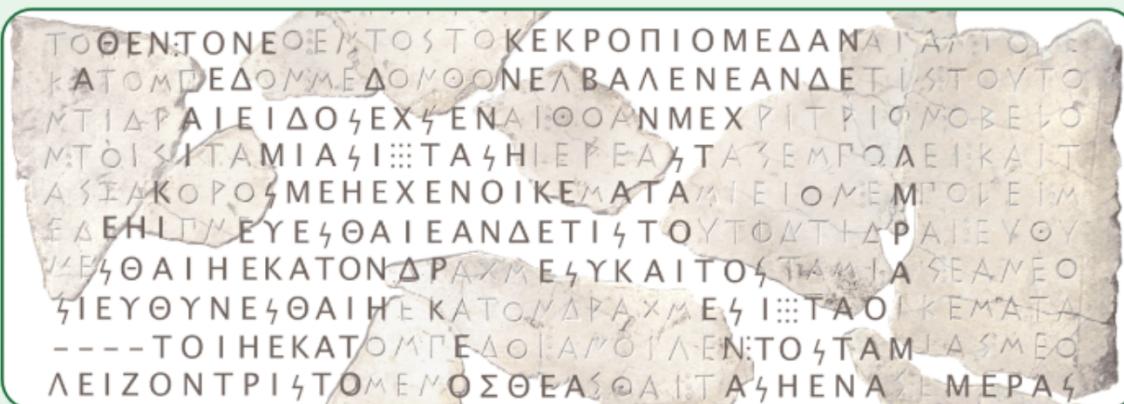
This image shows a restored decree concerning the Acropolis of Athens and dates 485/5 BCE.
(CC BY-SA 3.0, WikiMedia).

6. 跨学科性

以团队的方式对待科学，资助被忽视的角色，并促进竞争文化

许多最困难的科学问题都需要在不同领域的边界上取得进展。然而，当从业者聚集在一起时，例如在 Covid-19 期间，他们往往很难从多学科团队（每个人保留自己的学科角度）过渡到真正的跨学科团队，在跨学科团队中他们共同开发共享的想法和方法。这一挑战反映了科学知识日益专业化，以及拨款等激励措施，这些激励措施通常主要根据从业者的核心专业知识来评估他们。

默认情况下，人工智能在科学领域的努力通常是多学科的，但要想取得成功，它们需要成为真正的跨学科。首先是选择一个需要每种类型专业知识的问题，然后提供足够的时间和精力来培养围绕该问题的团队活力。例如，我们的伊萨卡项目利用人工智能对受损的古希腊铭文进行修复和归因，这可以帮助从业者研究过去文明的思想、语言和历史。为了取得成功，项目联合负责人雅尼斯·阿萨尔（Yannis Assael）必须加深对金石学的理解——对古代铭文的研究。鉴于直觉对其工作的重要性，该项目的金石学家也必须了解人工智能模型的工作原理。培养这些团队活力需要正确的激励措施。AlphaFold 2 突破的关键是让一个紧密的小团队专注于解决问题，而不是论文作者。这种类型的关注在工业实验室中更容易实现，但再次强调了与出版压力无关的长期公共研究经费的重要性。



该图片显示了一份关于雅典卫城的修复法令，日期为公元前 485/5 年。

(CC BY-SA 3.0, 维基媒体)。

To achieve genuine interdisciplinarity, organisations also need to create roles and career paths for individuals who can help blend disciplines. At Google DeepMind, our *research engineers* encourage a positive feedback loop between research and engineering, while our *programme managers* help to cultivate team dynamics within a research effort and create links across them. We also prioritise hiring individuals who enjoy finding and bridging connections between fields, as well as those that are motivated by rapidly upskilling in new areas. To encourage a cross-pollination of ideas, we also encourage scientists and engineers to regularly switch projects. Ultimately, the goal is to create a culture that encourages curiosity, humility and what the economic historian Joel Mokyr has referred to as '[contestability](#)' - where practitioners of all backgrounds feel empowered to present and constructively critique each other's work in open talks and discussion threads. ■

7. Adoption

Carefully consider the best access option and spotlight AI models' uncertainties

Many AI for Science models, such as AlphaFold or our weather forecasting work, are *specialised* in the sense that they perform a small number of tasks. But they are also *general* in the sense that [a large number of scientists are using them](#), for everything from understanding diseases to improving fishing programmes. This impact is [far from guaranteed](#). The [germ-theory of disease](#) took a long time to diffuse, while the downstream products that scientific breakthroughs could enable, such as [novel antibiotics](#), often lack the right market incentives.

When deciding how to release our models, we try to balance the desire for widespread adoption and validation from scientists with commercial goals and other considerations, such as potential safety risks. We also created a dedicated Impact Accelerator to drive adoption of breakthroughs and encourage socially beneficial applications that may not otherwise occur, including through partnerships with organisations like the [Drugs for Neglected Diseases Initiative](#), and the [Global Antibiotic Research & Development Partnership](#), that have a similar mandate.

To encourage scientists who could benefit from a new model or dataset to use it, [developers need to make it](#) as easy as possible for scientists to use and integrate into their workflows. With this in mind, for AlphaFold 2 we open-sourced the code but also partnered with [EMBL-EBI](#) to develop [a database](#) where scientists, including those with less computational skills and infrastructure, could search and download from a preexisting set of 200 million protein structures. AlphaFold 3 expanded the model's capabilities, leading to a combinatorial explosion in the number of potential predictions. This created a need for a new interface, the [AlphaFold Server](#), which allows scientists to create structures on-demand. The scientific community has also developed their own AlphaFold tools, such as [ColabFold](#), demonstrating the diversity of needs that exist, as well as the value of nurturing computational skills in the scientific community to address these needs.

为了实现真正的跨学科性，组织还需要为能够帮助融合学科的个人创造角色和职业道路。在 Google DeepMind，我们的研究工程师鼓励研究和工程之间建立积极的反馈循环，而我们的项目经理则帮助在研究工作中培养团队活力并在他们之间建立联系。我们还优先雇用那些喜欢在不同领域之间寻找和建立联系的人，以及那些有动力在新领域快速提高技能的人。为了鼓励思想的交叉传播，我们还鼓励科学家和工程师定期更换项目。最终的目标是创造一种鼓励好奇心、谦逊和经济历史学家 Joel Mokyr 所说的“可竞争性”的文化——在这种文化中，各种背景的从业者都感到有权在公开演讲和讨论中展示和建设性地批评彼此的工作。

7. 收养

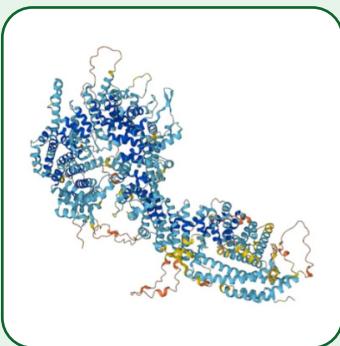
仔细考虑最佳访问选项并关注人工智能模型的不确定性

许多科学人工智能模型，例如 AlphaFold 或我们的天气预报工作，都是专门执行少量任务的。但它们也具有普遍性，因为大量科学家正在使用它们，从了解疾病到改善捕鱼计划等各个方面。这种影响远不能保证。疾病的细菌理论花了很多时间才传播开来，而科学突破可能带来的下游产品，如新型抗生素，往往缺乏适当的市场激励。

在决定如何发布我们的模型时，我们试图在科学家广泛采用和验证的愿望与商业目标和其他考虑因素（例如潜在的安全风险）之间取得平衡。我们还创建了一个专门的影响加速器，以推动突破性成果的采用，并鼓励原本可能不会发生的对社会有益的应用，包括通过与被忽视疾病药物倡议和全球抗生素研究与开发伙伴关系等具有类似使命的组织建立合作伙伴关系。

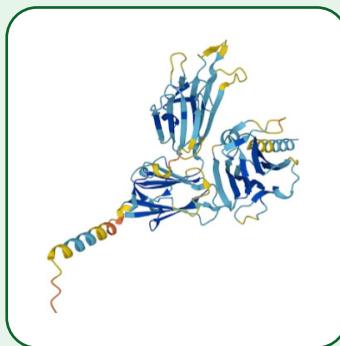
为了鼓励可以从新模型或数据集中受益的科学家使用它，开发人员需要让科学家尽可能轻松地使用和集成到他们的工作流程中。考虑到这一点，我们开源了 AlphaFold 2 的代码，同时还与 EMBL-EBI 合作开发了一个数据库，科学家（包括那些计算技能和基础设施较少的科学家）可以在其中搜索和下载预先存在的 2 亿个蛋白质结构集。AlphaFold 3 扩展了模型的功能，导致潜在预测数量的组合爆炸。这就产生了对新界面 AlphaFold Server 的需求，它允许科学家按需创建结构。科学界还开发了自己的 AlphaFold 工具，例如 ColabFold，展示了存在需求的多样性，以及培养科学界计算技能来满足这些需求的价值。

AlphaFold predictions referenced in publications



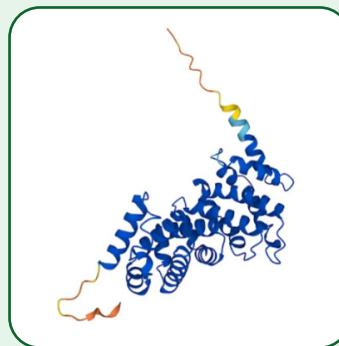
Nuclear pore complex protein Nup205

Part of a large complex that acts as a gateway in and out of the cell nucleus



Gametocyte surface protein P45/48

From the malaria parasite; a candidate protein for including in vaccines



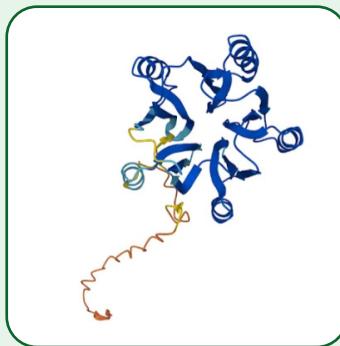
CCR4-NOT transcription complex subunit 9

Regulates an important cellular process (the rate of mRNA degradation)



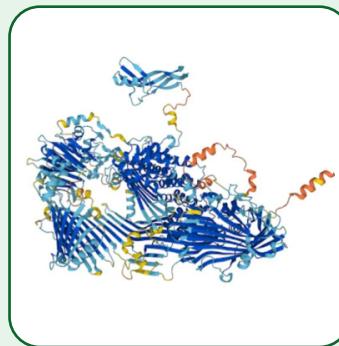
Ice nucleation protein

Bacterial protein that can trigger ice formation at relatively high temperatures, causing frost damage to plants



F20H23.2 protein

Plant protein; represents a potential new structural superfamily unlike anything seen before

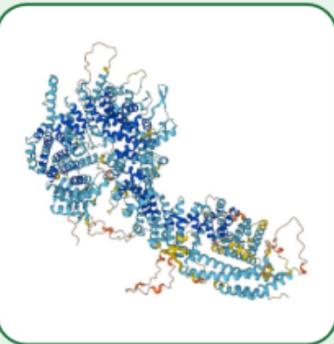


Vitellogenin

Involved in the immune system of egg-laying animals including honeybees

To date, more than 2 million users from over 190 countries have accessed the AlphaFold Protein Structure Database to view over 7 million structures.

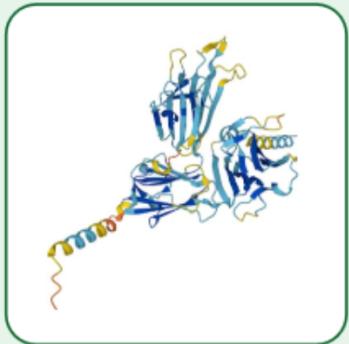
出版物中引用的 AlphaFold 预测



核孔复合蛋白Nup205

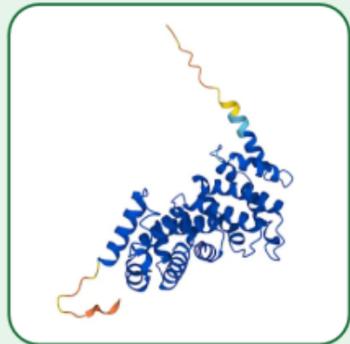
大型综合体的一部分，作为进出的门户

细胞核的



配子体表面
蛋白质 P45/48

来自马来寄生虫；疫苗中的候选蛋白质



CCR4-NOT 转录复合物亚基 9

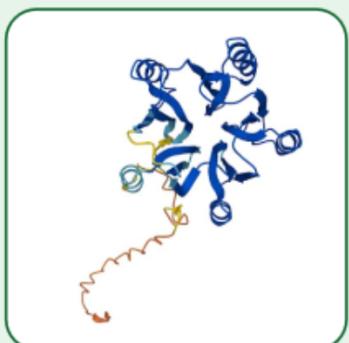
调节重要的细胞过程 (mRNA 降解率)



冰核蛋白

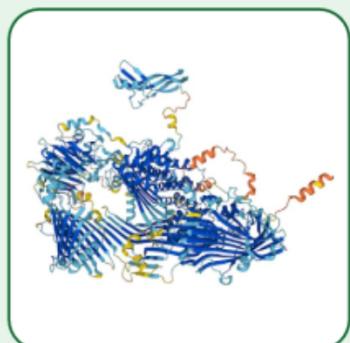
细菌蛋白可在相对较高的温度下引发结冰，从而导致霜冻

对植物的损害



F20H23.2蛋白

植物蛋白；代表了一个潜在的新结构超家族，与之前见过的任何东西都不一样



卵黄蛋白原

参与产卵动物（包括蜜蜂）的免疫系统

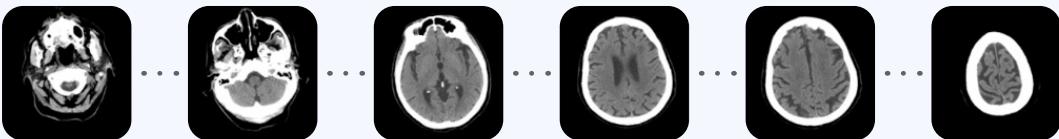
迄今为止，已有来自 190 多个国家的超过 200 万用户访问过

AlphaFold 蛋白质结构数据库可查看超过 700 万个结构。

Scientists also need to trust an AI model in order to use it. We expand on the reliability question below, but a useful starting point is to proactively signal how scientists should use a model, as well as its uncertainties. With AlphaFold, following dialogue with scientists the team developed uncertainty metrics that communicated how ‘confident’ the model was about a given protein structure prediction, supported by intuitive visualisations. We also partnered with EMBL-EBI to develop a [training module](#) that offered guidance on how to best use AlphaFold, including how to interpret the confidence metrics, supported by practical examples of how other scientists were using it. Similarly, our [Med-Gemini](#) system recently achieved state-of-the-art performance on answering health-related questions. It uses an *uncertainty-guided* approach that responds to a question by generating multiple ‘reasoning chains’ for how it might answer. It then uses the relative divergence between these initial answers to calculate how uncertain the answer is. Where uncertainty is high, it invokes web search to integrate the latest, up-to-date information. ■

Prompt

Given the image, what are the radiology findings?



Response

FINDINGS: mild cortical atrophy and periventricular white matter changes are noted compatible with microvasculopathy. there is proportionate ventriculomegaly. no bleed, mass effect, midline shift or abnormal intracranial fluid is seen. no evidence of acute intracranial hemorrhage, territorial infarct, abnormal fluid or mass effect.no abnormal extra-axial fluid collection is identified. no definitive evidence of intracranial hemorrhage or large territorial infarct within the limits of ct imaging. **old left basilar ganglia infarct.** no intracranial mass, swelling, hemorrhage, infarction or extra-axial collection is seen. visualized mastoid air cells: well pneumatized. visualized paranasal sinuses: well pneumatized. atherosclerotic calcification of vasculature at the skull base.

Radiologist Written Report

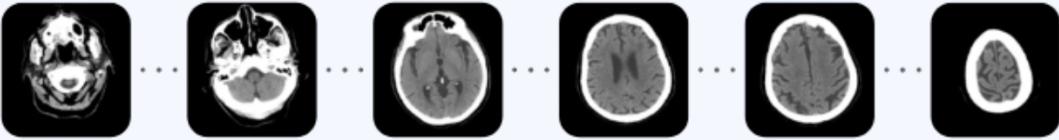
FINDINGS: the fourth ventricle is midline. the remaining ventricular system is within normal limits. there is mild chronic white matter microangiopathic disease and moderate generalized cortical atrophy. no bleed, mass effect, midline shift or abnormal intracranial fluid is seen. bone windows demonstrate an intact skull. the remaining visualized paranasal sinuses and mastoids are clear. chronic white matter microangiopathic disease, mild generalized cortical atrophy, not significantly changed since 2015. no evidence of acute intracranial hemorrhage, territorial infarct, abnormal fluid or mass effect.

Med-Gemini-3D is able to generate reports for CT scans, a far more complex form of imaging than standard X-rays. In this example, Med-Gemini-3D’s report has correctly included a pathology (highlighted in green) that was missed in the original radiologist’s report. Note that ‘basilar’ is a common mis-transcription of ‘basal’ that Med-Gemini has learned from the training data, though the meaning of the report is unchanged.

科学家还需要信任人工智能模型才能使用它。我们将在下面详细讨论可靠性问题，但一个有用的起点是主动表明科学家应该如何使用模型及其不确定性。在与科学家对话后，团队通过 AlphaFold 开发了不确定性指标，在直观可视化的支持下，传达模型对给定蛋白质结构预测的“信心”程度。我们还与 EMBL-EBI 合作开发了一个培训模块，为如何最好地使用 AlphaFold 提供指导，包括如何解释置信度指标，并辅以其他科学家如何使用它的实际示例。同样，我们的 MedGemini 系统最近在回答健康相关问题方面取得了最先进的性能。它使用不确定性引导的方法，通过生成多个“推理链”来回答问题，以了解如何回答。然后，它使用这些初始答案之间的相对差异来计算答案的不确定性。在不确定性较高的情况下，它会调用网络搜索来整合最新的信息。

迅速的

根据图像，放射学结果是什么？



回复

结果：轻度皮质萎缩和脑室周围白质变化与微血管病变相一致。存在成比例的脑室扩大。未见出血、占位效应、中线移位或颅内液异常。没有急性颅内出血、脑梗死、异常液体或占位效应的证据。没有发现异常的轴外液体收集。CT 成像范围内没有明确的颅内出血或大面积梗塞证据。陈旧性左侧基底神经节梗塞。未见内肿块、肿胀、出血、梗死或轴外聚集。可视化乳突气室：气动良好。

鼻旁窦可视化：气动良好。颅底脉管系统的动脉粥样硬化钙化。

放射科医生书面报告

结果：第四脑室位于中线。其余心室系统在正常范围内。存在轻度慢性白质微血管病和中度全身性皮质萎缩。未见出血、占位效应、中线移位或颅内液异常。骨窗展示了完整的头骨。其余可视化的鼻旁窦和乳突清晰可见。慢性白质微血管病变，轻度全身性皮质萎缩，自2015年以来无明显变化。无急性颅内出血、脑区梗死、异常液体或占位效应的证据。

Med-Gemini-3D 能够生成 CT 扫描报告，这是一种比标准 X 射线复杂得多的成像形式。在此示例中，Med-Gemini-3D 的报告正确包含了原始放射科医生报告中遗漏的病理学（以绿色突出显示）。请注意，“basilar”是 Med-Gemini 从训练数据中学到的“basal”的常见错误转录，尽管报告的含义没有变化。

8. Partnerships

Aim for early alignment and a clear value exchange

AI for Science efforts require a diversity of expertise that creates a strong need for partnership - both formal and informal - between public and private organisations. These partnerships are needed throughout the project lifecycle, from creating datasets to sharing the research. In particular, AI labs often need scientists to help evaluate an AI model's outputs. For example, [recent research](#) highlighted how senior material scientists with rich domain expertise were critical to evaluating whether the new materials predicted by an AI model were likely to be viable. Similarly, our protein design team partnered with research groups from the [Francis Crick Institute](#) to run wet lab experiments to test if our AI-designed proteins bound to their target and if this had the desired function, such as preventing SARS-CoV-2 from infecting cells. Our [FunSearch](#) method was able to develop a new construction for the [Cap Set](#) problem, which renowned mathematician Terence Tao once described as his [favourite open question](#). This was enabled by collaborating with Jordan Ellenberg, a professor of mathematics at the University of Wisconsin–Madison and a noted Cap Set expert. Given the central role played by industry labs in advancing AI capabilities, and the need for rich domain expertise, these public private partnerships will likely become increasingly important to advancing the AI for Science frontier and may require greater investment, such as more funding to support partnerships teams in universities and public research institutions.

Developing partnerships is difficult. When starting discussions, it is important to align early on the overall goal and address potentially thorny questions, such as what rights each party should have over the outputs, whether there should be a publication, whether the model or dataset should be open sourced, and what type of licensing should apply. Differences of opinion are natural and often reflect the incentives facing public and private organisations, which in turn vary greatly, depending on factors such as the maturity of the research or its commercial potential. The most successful partnerships involve a clear value-exchange that draws on the strengths of each organisation. For example, more than 2 million users from over 190 countries have used the [AlphaFold Protein Structure Database](#). This required a close collaboration to pair our AI model with the biocuration expertise and scientific networks of [EMBL-EBI](#). ■

8. 合作伙伴

旨在尽早协调和明确的价值交换

人工智能促进科学工作需要多样化的专业知识，这迫切需要公共和私人组织之间建立正式和非正式的伙伴关系。从创建数据集到共享研究成果，整个项目生命周期都需要这些合作伙伴关系。特别是，人工智能实验室通常需要科学家帮助评估人工智能模型的输出。例如，最近的研究强调了具有丰富领域专业知识的高级材料科学家对于评估人工智能模型预测的新材料是否可行至关重要。同样，我们的蛋白质设计团队与弗朗西斯·克里克研究所的研究小组合作，进行湿实验室实验，以测试我们的人工智能设计的蛋白质是否与其目标结合，以及是否具有所需的功能，例如防止 SARS-CoV-2 感染细胞。我们的 FunSearch 方法能够为 Cap Set 问题开发一种新的结构，著名数学家陶哲轩曾将其描述为他最喜欢的开放问题。这是通过与威斯康星大学麦迪逊分校数学教授、著名 Cap Set 专家 Jordan Ellenberg 的合作实现的。鉴于行业实验室在推进人工智能能力方面发挥的核心作用以及对丰富领域专业知识的需求，这些公私合作伙伴关系对于推进人工智能科学前沿可能变得越来越重要，并且可能需要更多投资，例如更多资金来支持合作伙伴关系大学和公共研究机构的团队。

发展伙伴关系是困难的。在开始讨论时，重要的是尽早调整总体目标并解决潜在的棘手问题，例如各方对输出应拥有哪些权利，是否应该发布，模型或数据集是否应该开源，以及应适用什么类型的许可。意见分歧是很自然的，通常反映了公共和私人组织面临的激励因素，而这些激励因素又存在很大差异，具体取决于研究的成熟度或其商业潜力等因素。最成功的合作伙伴关系涉及利用每个组织的优势进行明确的价值交换。例如，来自 190 多个国家的超过 200 万用户使用了 AlphaFold 蛋白质结构数据库。这需要密切合作，将我们的 AI 模型与 EMBL-EBI 的生物管理专业知识和科学网络相结合。

9. Safety & responsibility

Use assessments to explore trade-offs and inspire new types of evaluation methods

Scientists often disagree, sometimes strongly, about the potential benefits and risks that AI models may have on science, and on wider society. Conducting an ethics and safety assessment can help to frame the discussion and enable scientists to decide whether, and how, to develop a given AI model. A starting point is to identify the most important domains of impact, and to specify these domains at the right level of abstraction. There are increasingly sophisticated frameworks to [identify and categorise different AI risks](#), such as enabling mis- and disinformation. But these frameworks rarely consider the potential benefits of AI in the same domain, such as improving access to high-quality information synthesis, or the trade-offs that can occur, for example if you restrict access to an AI model or limit its capabilities. Assessments should also clarify their timescales, the relative certainty of any impact, and the relative importance, or *additionality*, of AI, to achieving it. For example, those worried about AI and climate change often focus on the immediate power needed to train large AI models, while AI proponents often focus on the less immediate, less clear, but potentially much larger downstream benefits to the climate from future AI applications. In carrying out their assessment, AI practitioners should also avoid over-indexing on the model's capabilities, which they will be closer to, and better understand the extent to which third parties will actually use it or be affected by it, which typically requires input from external experts to do well.

Practitioners also need new methods to better evaluate the potential risks and benefits of using AI in science. At present, many AI safety evaluations rely on specifying the types of content that a model should not output, and quantifying the extent to which the model adheres to this policy. These evaluations are useful for certain risks posed by using AI in science, such as generating inaccurate content. But for other risks, such as to biosecurity, the idea that we can reliably specify certain types of scientific knowledge as dangerous in advance has been [challenged](#), because of the dual-use nature of scientific knowledge, but also because such efforts tend to focus on what has caused harm historically, such as viruses from past outbreaks, rather than novel risks. A better approach may be to evaluate the [dangerous capabilities](#) of AI models, or the degree to which AI models provide an uplift to humans' dangerous capabilities. In many cases, these capabilities will also be dual-use, such as the ability to help design or execute experimental protocols. The degree to which these AI capabilities point to a risk, or an opportunity, will depend on how potential threat actors are assessed and how access to the model is governed. Beyond safety, evaluating other risks from using AI in science, such as to scientific creativity or reliability (which we discuss below), will require entirely new evaluation methods. Given the difficulty of researching and executing such evaluations, it makes sense to pursue them at the community-level, rather than each lab pursuing siloed efforts. ■

9. 安全与责任

使用评估来探索权衡并激发新型评估方法

对于人工智能模型对科学和更广泛的社会可能带来的潜在好处和风险，科学家们经常存在不同意见，有时甚至是强烈分歧。进行道德和安全评估有助于构建讨论，并使科学家能够决定是否以及如何开发特定的人工智能模型。起点是确定最重要的影响领域，并在正确的抽象级别指定这些领域。有越来越复杂的框架来识别和分类不同的人工智能风险，例如导致错误和虚假信息的风险。但这些框架很少考虑人工智能在同一领域的潜在好处，例如改善对高质量信息合成的访问，或可能发生的权衡，例如限制对人工智能模型的访问或限制其功能。评估还应阐明其时间表、任何影响的相对确定性，以及人工智能对于实现这一目标的相对重要性或额外性。例如，那些担心人工智能和气候变化的人往往关注训练大型人工智能模型所需的直接力量，而人工智能支持者则往往关注未来人工智能应用给气候带来的不太直接、不太明确但可能更大的下游效益。在进行评估时，人工智能从业者还应该避免过度索引模型的能力，这样他们会更接近模型的能力，并更好地了解第三方实际使用它或受其影响的程度，这通常需要来自外部专家做得很好。

从业者还需要新的方法来更好地评估在科学中使用人工智能的潜在风险和好处。目前，许多人工智能安全评估依赖于指定模型不应输出的内容类型，并量化模型遵守该政策的程度。这些评估对于在科学中使用人工智能带来的某些风险很有用，例如生成不准确的内容。但对于其他风险，例如生物安全，我们可以可靠地提前指定某些类型的科学知识为危险的想法受到了挑战，因为科学知识的双重用途性质，而且还因为此类努力往往集中于历史上造成伤害的因素，例如过去爆发的病毒，而不是新的风险。更好的方法可能是评估人工智能模型的危险能力，或者人工智能模型对人类危险能力的提升程度。在许多情况下，这些功能也将具有双重用途，例如帮助设计或执行实验方案的能力。这些人工智能功能指向风险或机遇的程度将取决于如何评估潜在威胁参与者以及如何管理对模型的访问。除了安全性之外，评估在科学中使用人工智能的其他风险，例如科学创造力或可靠性（我们将在下面讨论），将需要全新的评估方法。考虑到研究和执行此类评估的难度，在社区层面进行评估是有意义的，而不是每个实验室都进行孤立的努力。

Part C

The risks

C部分

风险

The risks

[Policy papers](#), [government documents](#) and [surveys of scientists](#) regularly cite certain risks from the growing use of AI in science. Three of these risks - to scientific creativity, reliability, and understanding - mainly relate to how science is practised. Two other risks - to equity and the environment - mainly relate to how science represents and affects wider society. The use of AI is often presented exclusively as a risk to these domains, and the domains, such as scientific reliability, or the environment, are often portrayed in stable, somewhat idealised terms, that can overlook the wider challenges that they face. We believe that using AI in science will ultimately benefit each of these five domains, because there are opportunities to mitigate the risks that AI poses, and to use AI to help address wider challenges in these areas, in some cases profoundly. Achieving a beneficial outcome will likely be harder for inequity, which is ingrained into AI and science at multiple levels, from the make-up of the workforce to the data underpinning research, and for scientific creativity, which is highly subjective and so individuals may reasonably disagree about whether a certain outcome is positive. These nuances increase the value of scientists, policymakers and others articulating their expectations for how using AI in science will affect each of these 5 areas. ■

5 risks of using AI to accelerate science



1. Creativity



2. Reliability



3. Understanding



4. Equity



5. Environment

风险

政策文件、政府文件和科学家调查经常提到人工智能在科学领域日益广泛使用所带来的某些风险。其中三个风险——科学创造力、可靠性和理解——主要与科学的实践方式有关。另外两个风险——公平和环境——主要与科学如何代表和影响更广泛的社会有关。人工智能的使用通常被完全视为这些领域的风险，而科学可靠性或环境等领域通常用稳定的、有些理想化的术语来描述，这可能会忽视它们面临的更广泛的挑战。我们相信，在科学中使用人工智能最终将使这五个领域受益，因为有机会减轻人工智能带来的风险，并利用人工智能帮助解决这些领域更广泛的挑战，在某些情况下意义深远。对于不平等来说，实现有益的结果可能会更加困难，不平等在人工智能和科学的多个层面上根深蒂固，从劳动力的构成到支持研究的数据，以及科学创造力，这是高度主观的，因此个人可以合理地对于某个结果是否积极存在分歧。这些细微差别增加了科学家、政策制定者和其他人表达他们对在科学中使用人工智能将如何影响这 5 个领域的期望的价值。

使用人工智能加速科学发展的 5 个风险



1. 创造力



2. 可靠性



3. 理解



4. 股权



5. 环境

1. Creativity

Will AI lead to less novel, counterintuitive, breakthroughs?

Scientific creativity describes the creation of something new that is useful. In practice, the extent to which a scientist views a new idea, method, or output as creative typically rests on more subjective factors, such as its perceived simplicity, counterintuitiveness, or beauty. Today, scientific creativity is undermined by the relative homogeneity of the scientific workforce, which narrows the diversity of ideas. The pressure on researchers to 'publish or perish' also incentivises 'crowd-following publications on less risky topics', rather than the kind of deep work, or bridging of concepts across disciplines, that often underpins creative breakthroughs. This may explain why the share of disruptive scientific ideas that cause a field to veer off into a new direction seems to be declining, beyond what may be normally expected, as science expands.

Some scientists worry that using AI may exacerbate these trends, by undermining the more intuitive, unorthodox, and serendipitous approaches of human scientists, such as Galileo's hypothesis that the earth rotates on its axis. This could happen in different ways. One concern is that AI models are trained to minimise anomalies in their training data, whereas scientists often amplify anomalies by following their intuitions about a perplexing data point. Others worry that AI systems are trained to perform specific tasks, and so relying on them will forgo more serendipitous breakthroughs, such as researchers unexpectedly finding solutions to problems that they weren't studying. At the community level, some worry that if scientists embrace AI en masse, it may lead to a gradual homogenisation of outputs, for example if LLMs produce similar suggestions in response to the queries of different scientists. Or if scientists over-focus on disciplines and problems that are best-suited to AI.

Maintaining support for exploratory research and non-AI research could help to mitigate some of these risks. Scientists could also tailor how they use AI so that it boosts rather than detracts from their own creativity, for example by fine-tuning LLMs to suggest more personalised research ideas, or to help scientists better elicit their own ideas, similar to our early efforts to develop AI tutors that could help students to better reflect on a problem, rather than just outputting answers to questions. AI could also enable new types of scientific creativity that may be unlikely to otherwise occur. One type of AI creativity is *interpolation* where AI systems identify novel ideas within their training data, particularly where humans' ability to do this is limited, such as efforts to use AI to detect anomalies in massive datasets from Large Hadron Collider experiments. A second type is *extrapolation*, where AI models generalise to more novel solutions outside their training data, such as the famous move 37 that our AlphaGo system came up with, that stunned human Go experts, or the novel maths proofs and non-obvious constructions that our AlphaProof and AlphaGeometry 2 systems produced. A third type is *invention*, where AI systems come up with an entirely new theory or scientific system, completely removed from its training data, akin to the original development of general relativity, or the creation of complex numbers. AI systems do not currently demonstrate such creativity, but new approaches could potentially unlock this, such as multi-agent systems that are optimised for different goals, like novelty and counterintuitiveness, or AI models that are trained to generate novel scientific problems in order to inspire novel solutions. ■

1. 创造力

环境人工智能会带来不那么新颖、违反直觉的突破吗？

科学创造力描述了有用的新事物的创造。在实践中，科学家将新想法、方法或成果视为创造性的程度通常取决于更主观的因素，例如其感知的简单性、反直觉或美感。如今，科学劳动力的相对同质性削弱了科学创造力，从而缩小了思想的多样性。研究人员“要么发表，要么灭亡”的压力也激励他们在风险较小的主题上“大众追随”出版物，而不是那种经常支撑创造性突破的深度工作或跨学科概念的桥梁。这或许可以解释为什么随着科学的发展，导致某个领域转向新方向的颠覆性科学思想的份额似乎正在下降，超出了通常的预期。

一些科学家担心，人工智能的使用可能会加剧这些趋势，因为它会破坏人类科学家更直观、非正统和偶然的方法，例如伽利略关于地球绕轴旋转的假设。这可能以不同的方式发生。一个令人担忧的问题是，人工智能模型经过训练，可以最大限度地减少训练数据中的异常情况，而科学家经常通过遵循他们对令人困惑的数据点的直觉来放大异常情况。其他人担心人工智能系统经过训练可以执行特定任务，因此依赖它们会放弃更多偶然的突破，例如研究人员意外地找到他们没有研究的问题的解决方案。在社区层面，一些人担心，如果科学家集体拥抱人工智能，可能会导致输出逐渐同质化，例如，如果LLMs针对不同科学家的询问提出类似的建议。或者，如果科学家过度关注最适合人工智能的学科和问题。

维持对探索性研究和非人工智能研究的支持可能有助于减轻其中一些风险。科学家还可以调整他们使用人工智能的方式，使其增强而不是削弱他们自己的创造力，例如通过微调LLMs来提出更个性化的研究想法，或者帮助科学家更好地引出他们自己的想法，类似于我们早期开发人工智能导师的努力，可以帮助学生更好地思考问题，而不仅仅是输出问题的答案。人工智能还可以实现新型的科学创造力，否则这些创造力可能不太可能发生。人工智能创造力的一种类型是插值，人工智能系统在训练数据中识别新颖的想法，特别是在人类执行此操作的能力有限的情况下，例如使用人工智能来检测大型强子对撞机实验中的大量数据集中的异常。第二种类型是外推法，人工智能模型可以推广到训练数据之外的更新颖的解决方案，例如我们的AlphaGo系统提出的著名的第37步，令人类围棋专家震惊，或者新颖的数学证明和非显而易见的结构我们生产的AlphaProof和AlphaGeometry 2系统。第三种类型是发明，其中人工智能系统提出了一种全新的理论或科学系统，完全从其训练数据中删除，类似于广义相对论的原始发展或复数的创建。人工智能系统目前尚未表现出这种创造力，但新方法可能会解锁这一点，例如针对不同目标（例如新颖性和反直觉性）进行优化的多智能体系统，或者经过训练以产生新颖的科学问题以激发灵感的人工智能模型新颖的解决方案。

2. Reliability

Will AI make science less self-correcting?

Reliability describes the ability of scientists to depend upon each others' findings, and trust that they are not due to chance or error. Today, a series of [interrelated challenges](#) weaken the reliability of science, including the p-hacking and publication bias which can lead [researchers to underreport negative results; a lack of standardisation](#) in how scientists carry out routine scientific tasks; mistakes, [for example in how scientists use statistical methods; scientific fraud; and challenges with the peer review process, including a lack of qualified peer reviewers.](#)

Some scientists worry that AI will exacerbate these challenges as some AI research also features [bad practices](#), such as practitioners cherrypicking the evaluations they use to assess their models' performance. AI models, particularly LLMs, are also prone to 'hallucinate' outputs, including scientific citations, that are false or misleading. Others worry that LLMs may lead to a flood of low-quality papers similar to those that ['paper mills' churn out](#). The community is working on mitigations to these problems, including [good practice checklists](#) for researchers to adhere to and different types of AI factuality research, such as training AI models to [ground their outputs to trusted sources](#), or to help [verify the outputs](#) of other AI models.

Scientists could also potentially use AI to improve the reliability of the wider research base. For instance, if AI can help to automate aspects of data annotation or experiment design, this could provide much-needed standardisation in these areas. As AI models get better at grounding their outputs to citations, they could also help scientists and policymakers do more [systematic reviews](#) of the evidence base, for example in climate change, where groups like the Intergovernmental Panel on Climate Change are already struggling to keep up with the [inexorable rise in publications](#). Practitioners could also use AI to help detect mistaken or fraudulent images, or misleading scientific claims, as seen in the [recent trial](#) by the Science journals of an AI image analysis tool. More speculatively, AI could potentially help with aspects of peer review, given that some scientists already use LLMs to help critique their own papers, and to help validate the outputs of AI models, [for example in theorem proving](#). However, there are also reasonable concerns about confidentiality, the ability of AI systems to detect truly novel work, and the need for buy-in from scientists given the consequential role that peer review plays in processes such as grant approvals. ■

2. 可靠性

人工智能会让科学减少自我纠正吗？

可靠性描述了科学家相互依赖研究结果的能力，并相信这些结果不是偶然或错误造成的。如今，一系列相互关联的挑战削弱了科学的可靠性，包括 p-hacking 和出版偏见，这可能导致研究人员少报负面结果；科学家如何执行日常科学任务缺乏标准化；错误，例如科学家如何使用统计方法；科学欺诈；同行评审过程面临的挑战，包括缺乏合格的同行评审员。

一些科学家担心人工智能会加剧这些挑战，因为一些人工智能研究也存在不良做法，例如从业者挑选用于评估模型性能的评估。人工智能模型，特别是LLMs，也容易产生虚假或误导性的“幻觉”输出，包括科学引文。其他人担心LLMs可能会导致大量低质量论文，类似于“造纸厂”生产的论文。社区正在努力缓解这些问题，包括供研究人员遵守的良好实践清单以及不同类型的人工智能事实性研究，例如训练人工智能模型以将其输出建立在可信来源上，或帮助验证其他人工智能模型的输出。

科学家们还可以利用人工智能来提高更广泛研究基础的可靠性。例如，如果人工智能可以帮助实现数据注释或实验设计方面的自动化，那么这可以提供这些领域急需的标准化。随着人工智能模型能够更好地将其输出结果与引文联系起来，它们还可以帮助科学家和政策制定者对证据基础进行更系统的审查，例如在气候变化方面，政府间气候变化专门委员会等组织已经在努力跟上气候变化的步伐。出版物的不可阻挡的增长从业者还可以使用人工智能来帮助检测错误或欺诈的图像，或误导性的科学主张，正如《科学》杂志最近对人工智能图像分析工具的试验所示。更具推测性的是，鉴于一些科学家已经使用 LLMs 来帮助批评自己的论文，并帮助验证人工智能模型的输出（例如在定理证明中），人工智能可能会在同行评审方面提供帮助。然而，考虑到同行评审在拨款批准等过程中发挥的重要作用，人们对保密性、人工智能系统检测真正新颖的工作的能力以及科学家的支持的必要性存在合理的担忧。

3. Understanding

Will AI lead to useful predictions at the expense of deeper scientific understanding?

In a recent [Nature survey](#), scientists cited a reliance on pattern matching at the expense of deeper understanding as the biggest risk from using AI in science. Understanding is not always necessary to discover new scientific phenomena, such as superconductivity, or to develop useful applications, such as drugs. But most scientists view understanding as one of their primary goals, as the deepest form of human knowledge. Concerns about AI undermining scientific understanding include the argument that modern deep learning methods are *atheoretical* and do not incorporate or contribute to theories for the phenomena that they predict. Scientists also worry that AI models are *uninterpretable*, in the sense that they are not based on clear sets of equations and parameters. There is also a concern that any explanation for an AI model's outputs will not be [accessible or useful](#) to scientists. Taken together, AI models may provide useful predictions about the structure of a protein, or the weather, but will they be able to help scientists understand why a protein folds a certain way, or how atmospheric dynamics lead to weather shifts?

Concerns about replacing '*real, theoretical science*' with '*low-brow ... computation*' are not new and [were levelled](#) at past techniques, such as the Monte Carlo method. Fields that merge engineering and science, such as [synthetic biology](#), have also faced accusations of prioritising useful applications over deeper scientific understanding. Those methods and technologies led to gains in scientific understanding and we are confident that AI will too, even if some of these gains will be hard to predict in advance. First, [most AI models are not atheoretical](#) but build on prior knowledge in different ways, such as in the construction of their datasets and evaluations. Some AI models also have interpretable outputs. For example, our [FunSearch](#) method outputs computer code that also describes how it arrived at its solution.

Researchers are also working on explainability techniques that could shed light on how AI systems work, such as [efforts to identify the 'concepts' that a model learns](#). Many of these explainability techniques have [important limitations](#), but they have already enabled scientists to extract new scientific hypotheses from AI models. For example, transcription factors are proteins that bind to DNA sequences to activate or repress the expression of a nearby gene. One [AI research effort](#) was able to predict the relative contribution of each base in a DNA sequence to the binding of different transcription factors and to explain this result using concepts familiar to biologists. A bigger opportunity may be to learn entirely new concepts based on how AI systems learn. For example, our researchers recently [demonstrated](#) that our AlphaZero system learned 'superhuman' knowledge about playing chess, including unconventional moves and strategies, and used another AI system to extract these concepts and teach them to human chess experts.

3. 理解

人工智能是否会以牺牲更深入的科学理解为代价来带来有用的预测？

在《自然》杂志最近的一项调查中，科学家们指出，在科学中使用人工智能的最大风险是依赖模式匹配而牺牲更深入的理解。理解并不总是发现新的科学现象（例如超导性）或开发有用的应用（例如药物）所必需的。但大多数科学家将理解视为他们的主要目标之一，视为人类知识的最深刻形式。对人工智能破坏科学理解的担忧包括这样的论点：现代深度学习方法是非理论性的，不会纳入或贡献于它们所预测的现象的理论。科学家还担心人工智能模型无法解释，因为它们不是基于清晰的方程和参数集。还有人担心，科学家无法获取或使用对人工智能模型输出的任何解释。总而言之，人工智能模型可能会提供有关蛋白质结构或天气的有用预测，但它们是否能够帮助科学家理解为什么蛋白质以某种方式折叠，或者大气力学如何导致天气变化？

用“低俗计算”取代“真实的理论科学”的担忧并不新鲜，而且

与过去的技术（例如蒙特卡罗方法）持平。合成生物学等融合工程和科学的领域也面临着优先考虑有用应用而不是更深入的科学理解的指责。这些方法和技术带来了科学理解的进步，我们相信人工智能也会如此，即使其中一些进步很难提前预测。首先，大多数人工智能模型并不是非理论性的，而是以不同的方式建立在先验知识的基础上，例如构建数据集和评估。一些人工智能模型还具有可解释的输出。例如，我们的 FunSearch 方法输出计算机代码，该代码还描述了它如何得出解决方案。

研究人员还在研究可解释性技术，这些技术可以揭示人工智能系统的工作原理，例如识别模型学习的“概念”的努力。许多可解释性技术都有重要的局限性，但它们已经使科学家能够从人工智能模型中提取新的科学假设。例如，转录因子是与 DNA 序列结合以激活或抑制附近基因表达的蛋白质。一项人工智能研究工作能够预测 DNA 序列中每个碱基对不同转录因子结合的相对贡献，并使用生物学家熟悉的概念解释这一结果。更大的机会可能是根据人工智能系统的学习方式来学习全新的概念。例如，我们的研究人员最近证明，我们的 AlphaZero 系统学习了有关下棋的“超人”知识，包括非常规的走法和策略，并使用另一个人工智能系统提取这些概念并将其教授给人类国际象棋专家。

Even without explainability techniques, AI will improve scientific understanding simply by opening up new research directions that would otherwise be prohibitive. For example, by unlocking the ability to generate a huge number of synthetic protein structures, AlphaFold enabled scientists to search across *protein structures*, rather than just across protein sequences. [One group](#) used this approach to discover an ancient member of the Cas13 protein family that offers promise for editing RNA, including to help diagnose and treat diseases. This discovery also challenged previous assumptions about how Cas13 evolved. Conversely, [efforts to modify](#) the AlphaFold model architecture to incorporate more prior knowledge led to worse performance. This highlights the trade-off that can occur between accuracy and interpretability, but also how AI systems could advance scientific understanding not *in spite of* their opacity, but *because of* it, as this opacity can stem from their ability to operate in high-dimensional spaces that may be uninterpretable to humans, but necessary to making scientific breakthroughs. ■

4. Equity

Will AI make science less representative, and useful, to marginalised groups?

Inequity is starkly visible in the scientific workforce, in the questions they study, in the data and models they develop, and in the benefits and harms that result. These inequities are related and can compound over time. For example, a small number of labs and individuals in higher-income cities account for a [disproportionate share of scientific outputs](#). [Studies to identify genetic variants associated with disease](#) rely heavily on data from European ancestry groups, while the [neglected tropical diseases](#) that disproportionately affect poor countries receive relatively little research funding. In agriculture, crop innovations focus on pests that are most common in high-income countries, and are then [inappropriately used on different pests in lower-income countries](#), hurting yields. Despite positive trends, women [account](#) for just 33% of scientists and have [long been underrepresented in clinical trials, particularly women of colour](#).

Observers worry that the growing use of AI in science could exacerbate these inequities. AI and computer science workforces are less representative, in terms of [gender, ethnicity and the location of leading labs](#), than many other scientific disciplines and so AI's growing use could hurt broader representation in science. As a data-driven technology, AI also risks inheriting and entrenching the biases found in scientific datasets.

There are also opportunities to use AI to reduce inequities in science, albeit not in lieu of more systemic change. If AI models are provided via low-cost servers or databases, they could make it easier and cheaper for scientists, including those from underrepresented groups, to study traditionally neglected problems, similar to how [releasing more satellite data led to more research from underrepresented communities](#). By ingesting more data, AI models may also be able to learn more universal patterns about the complex systems that scientists study, making these models more robust and less prone to biases. For example, owing to their non-representative data, studies that identify genetic variants associated with disease can pick up *confounding*, rather than *causal* variants. Conversely, some early attempts to train [AI models](#) on larger datasets of protein structures and genetic variants, including data across species, perform better at predicting individuals at the greatest risk for disease, with fewer discrepancies across population groups. Ultimately, however, improving equity will require long-term efforts, such as the [H3Africa](#) initiative in genomics and the [Deep Learning Indaba](#) initiative for AI, that aim to build up scientific infrastructure, communities, and education where it is most lacking. ■

即使没有可解释性技术，人工智能也将通过开辟新的研究方向来提高科学理解，否则这些研究方向将令人望而却步。例如，通过解锁生成大量合成蛋白质结构的能力，AlphaFold 使科学家能够搜索蛋白质结构，而不仅仅是蛋白质序列。一个研究小组利用这种方法发现了 Cas13 蛋白家族的一个古老成员，它为编辑 RNA 提供了希望，包括帮助诊断和治疗疾病。这一发现也挑战了之前关于 Cas13 如何进化的假设。相反，修改 AlphaFold 模型架构以纳入更多先验知识的努力

知识导致表现更差。这凸显了准确性和可解释性之间可能发生的权衡，也凸显了人工智能系统如何能够推进科学理解，尽管它们不透明，但正因为如此，因为这种不透明性可能源于它们在高维空间中运行的能力这对于人类来说可能是无法解释的，但却是实现科学突破所必需的。

4、股权

人工智能是否会降低科学对边缘群体的代表性和实用性？

在科学工作者中、在他们研究的问题中、在他们开发的数据和模型中、以及由此产生的利益和危害中，不平等现象是显而易见的。这些不平等现象是相互关联的，并且会随着时间的推移而加剧。例如，高收入城市的少数实验室和个人在科学产出中占据了不成比例的份额。识别与疾病相关的遗传变异的研究在很大程度上依赖于欧洲血统群体的数据，而对贫穷国家造成不成比例影响的被忽视的热带疾病却获得相对较少的研究经费。在农业领域，作物创新主要针对高收入国家最常见的害虫，然后不适当用于低收入国家的不同害虫，从而损害了产量。尽管趋势积极，但女性科学家仅占 33%

长期以来，临床试验中的代表性不足，尤其是有色人种女性。

观察人士担心，人工智能在科学领域的日益普及可能会加剧这些不平等。与许多其他科学学科相比，人工智能和计算机科学劳动力在性别、种族和领先实验室位置方面的代表性较低，因此人工智能的日益增长的使用可能会损害科学领域更广泛的代表性。作为一种数据驱动的技术，人工智能还面临继承和巩固科学数据集中发现的偏见的风险。

也有机会利用人工智能来减少科学领域的不平等，尽管不能取代更系统性的变革。如果通过低成本服务器或数据库提供人工智能模型，它们可以使科学家（包括来自代表性不足群体的科学家）更容易、更便宜地研究传统上被忽视的问题，类似于发布更多卫星数据如何导致代表性不足社区进行更多研究。通过摄取更多数据，人工智能模型还可能能够了解科学家研究的复杂系统的更多通用模式，从而使这些模型更加稳健且不易出现偏差。例如，由于数据不具有代表性，识别与疾病相关的遗传变异的研究可能会发现混杂变异，而不是因果变异。相反，一些早期尝试在更大的蛋白质结构和遗传变异数据集（包括跨物种数据）上训练人工智能模型，在预测患病风险最大的个体方面表现更好，并且不同人群之间的差异较小。然而，最终，改善公平性需要长期努力，例如基因组学领域的 H3Africa 计划和人工智能深度学习 Indaba 计划，这些计划旨在最缺乏的地方建立科学基础设施、社区和教育。

5. The environment

Will AI hurt or help efforts to achieve NetZero?

Given their desire to understand the natural world, many scientists have long been active in efforts to protect the environment, from providing [early evidence](#) about climate change to developing photovoltaic cells. In recent years, a growing number of scientists have [voiced concerns](#) about the potential impact of AI on the environment and developed [methodologies](#) to try to quantify these impacts. Most concerns focus on the [potential impact](#) of training and using LLMs on greenhouse gas emissions, as well as related concerns, such as about the [water needed to cool data centres](#). One way to think about these effects is the [life cycle](#) approach, which captures both direct and indirect effects. Direct effects include the emissions from [building](#) and powering the data centres and devices that AI models are trained and run on. There is no comprehensive estimate for *all* direct emissions from AI. However, a 2021 [estimate](#) suggested that cloud and hyperscale data centres, where many large AI models are trained and deployed, accounted for just 0.1-0.2% of global emissions.

As the size of LLMs continues to grow, observers have [cautioned](#) that these figures may increase, potentially significantly. However, many users of LLMs, including scientists, will be able to fine-tune them, or use their predictions, at a relatively low compute cost, rather than training them from scratch. Efforts are also underway to make LLMs more efficient, and the [history of digital technology suggests](#) that sizable gains are possible, not least due to the commercial pressures to deliver faster and cheaper AI models. In some instances, the emissions from AI models will be lower than other approaches. For example, our internal analysis suggests that determining the structures of a small number (<10) of proteins experimentally uses roughly the same energy as a full training run of AlphaFold 2. These results need to be interpreted carefully, as AI simulations rely on, and inform, physical experiments, rather than substituting for them. But they also show how AI could enable a larger amount of scientific activity at a lower average energy cost.

Crucially, the *direct* effects of AI on emissions, whether positive or negative, will likely be minor compared to the *indirect* effects that AI-enabled applications have on emissions. Using AI in science opens up three major opportunities to reduce [emissions](#). First, progress at the nexus between AI, maths and computer science could dramatically improve the efficiency of the Internet, from [designing more efficient chips](#) to [finding more efficient algorithms](#) for routine tasks. As a growing share of the economy moves online, this should help to offset emissions across these sectors. AI could accelerate the development and use of renewable energy, for example by [designing new materials](#), such as for batteries or solar panels, by [optimising how the grid operates](#) and how it [integrates renewables](#), and via more transformative but uncertain opportunities like fusion. Finally, the world is [already getting warmer](#) and AI could help to better prepare for extreme weather events. For example, our weather forecasting model recently [correctly predicted](#), seven days in advance, that the deadly [Hurricane Beryl](#) would make ‘landfall’ in Texas. Non-AI models had originally predicted landfall in Mexico before correcting their prediction to Texas three days before it occurred. ■

5、环境

人工智能会损害还是有助于实现 NetZero 的努力？

鉴于他们渴望了解自然世界，许多科学家长期以来一直积极致力于保护环境，从提供有关气候变化的早期证据到开发光伏电池。近年来，越来越多的科学家对人工智能对环境的潜在影响表示担忧，并开发了方法来尝试量化这些影响。大多数担忧集中在培训和使用LLMs对温室气体排放的潜在影响，以及相关的担忧，例如冷却数据中心所需的水。考虑这些影响的一种方法是生命周期方法，它捕获直接和间接影响。直接影响包括人工智能模型训练和运行的数据中心和设备的建设和供电所产生的排放。目前还没有对人工智能的所有直接排放进行全面的估计。然而，2021 年的估计表明，训练和部署许多大型人工智能模型的云和超大规模数据中心仅占全球排放量的 0.1-0.2%。

随着LLMs的规模持续增长，观察家警告说，这些数字可能会增加，而且可能会显着增加。然而，LLMs 的许多用户（包括科学家）将能够以相对较低的计算成本对其进行微调或使用他们的预测，而不是从头开始训练它们。人们也在努力提高LLMs的效率，数字技术的历史表明，巨大的收益是可能的，尤其是由于提供更快、更便宜的人工智能模型的商业压力。在某些情况下，人工智能模型的排放量将低于其他方法。例如，我们的内部分析表明，通过实验确定少量（<10）蛋白质的结构所使用的能量与 AlphaFold 2 的完整训练运行大致相同。这些结果需要仔细解释，因为人工智能模拟依赖于，并为物理实验提供信息，而不是替代它们。但它们也展示了人工智能如何以较低的平均能源成本实现更大量的科学活动。

至关重要的是，与人工智能应用对排放的间接影响相比，人工智能对排放的直接影响（无论是积极还是消极）可能很小。在科学中使用人工智能为减少排放开辟了三大机遇。首先，人工智能、数学和计算机科学之间联系的进步可以极大地提高互联网的效率，从设计更高效的芯片到为日常任务寻找更高效的算法。随着越来越多的经济体转向线上，这应该有助于抵消这些行业的排放。人工智能可以加速可再生能源的开发和使用，例如通过设计电池或太阳能电池板等新材料，优化电网的运行方式及其整合可再生能源的方式，以及通过聚变等更具变革性但不确定的机会。最后，世界已经变暖，人工智能可以帮助更好地应对极端天气事件。例如，我们的天气预报模型最近提前 7 天正确预测了致命的飓风

贝丽尔将在德克萨斯州“登陆”。非人工智能模型最初预测登陆墨西哥，然后在登陆三天前将其预测修正到德克萨斯州。

Part D

The policy response

D部分

政策

回复

The policy response

Given the importance of scientific progress to almost every major economic, environmental and security goal, it follows that science, and the potential for AI to accelerate it, [should be a top priority for any government](#). What should a new AI for Science policy agenda look like? Policymakers can start by implementing the many good science and innovation policy ideas that [already exist](#) and which make even more sense in an era of AI-enabled science. For example, AI will improve the return on science research funding and so it provides a strong rationale to invest more in it and to trial new ideas to [speed up](#) and [experiment with](#) how this funding is allocated. On compute, governments could implement the idea set out in the UK's [Independent Review](#) to empower a dedicated body to continually assess and advise governments on potential investments. To support AI for Science startups, policymakers can improve their [spin-out policies](#) and support well-run [start-up incubators](#) and [fellowships](#).

But ambitious new policies are also needed to capitalise on the AI for Science opportunity. We share **four ideas** below. They are intended to be widely applicable, though the precise details would need to be tailored to the specific context of a country, taking into account national priorities, unique strengths and the institutional landscape. ■

1. Define the ‘Hilbert Problems’ for AI in Science

Scientific progress rests on picking the right problems. In 1900, the German mathematician David Hilbert [published 23 unsolved problems](#) that proved hugely influential for the subsequent direction of 20th century mathematics. As part of upcoming international events such as the [AI Action Summit](#) in Paris, policymakers, AI labs and science funders could launch a public call for scientists and technologists to identify the most important AI-shaped scientific problems, backed by a major new global fund to drive progress on them. Submissions should specify why the problem is important, why it is suited to modern AI systems, why it may be otherwise neglected, the data bottlenecks that exist, and how near-term technical progress could be evaluated.

The top ideas could form the basis of new scientific competitions, where scientists compete to solve these problems with AI, supported by new datasets, evaluation methods and competitive benchmarks. These could build on the [recent flurry of competitions that have emerged to evaluate the scientific capabilities of AI models](#), and include a new [AI for Science Olympiad](#) to attract exceptional young talent from across the world to the field. Beyond its direct impacts, the AI for Science ‘Hilbert Problems’ initiative could provide a welcome focal point for international scientific collaboration and funding, and inspire a new generation of interdisciplinary scientists to identify and pursue AI-shaped problems. ■

政策回应

鉴于科学进步对几乎所有主要经济、环境和安全目标的重要性，科学以及人工智能加速科学进步的潜力应该成为任何政府的首要任务。新的人工智能科学政策议程应该是什么样的？政策制定者可以首先实施许多已经存在的良好科学和创新政策理念，这些理念在人工智能支持的科学时代更有意义。例如，人工智能将提高科学研究经费的回报，因此它提供了强有力的理由来对其进行更多投资，并尝试新的想法，以加快和试验如何分配这些资金。在计算方面，政府可以实施英国独立审查中提出的想法，授权专门机构持续评估政府的潜在投资并为其提供建议。为了支持科学初创企业的人工智能，政策制定者可以改进其分拆政策并支持运行良好的初创企业孵化器和奖学金。

但还需要雄心勃勃的新政策来利用人工智能促进科学的机会。我们分享以下四个想法。它们旨在广泛适用，但具体细节需要根据国家的具体情况进行调整，同时考虑到国家优先事项、独特优势和制度环境。

1. 定义科学中人工智能的“希尔伯特问题”

科学进步取决于选择正确的问题。1900年，德国数学家大卫·希尔伯特发表了23个未解决的问题，这些问题对20世纪数学的后续方向产生了巨大影响。作为即将在巴黎举行的人工智能行动峰会等国际活动的一部分，政策制定者、人工智能实验室和科学资助者可以公开呼吁科学家和技术人员确定最重要的人工智能型科学问题，并得到一个新的全球重大基金的支持，以解决人工智能问题。推动他们取得进步。提交的材料应具体说明为什么该问题很重要，为什么它适合现代人工智能系统，为什么它可能会被忽视，存在的数据瓶颈，以及如何评估近期技术进展。

顶尖的想法可以构成新的科学竞赛的基础，科学家们在新的数据集、评估方法和竞争基准的支持下，竞相利用人工智能解决这些问题。这些可以建立在最近出现的一系列评估人工智能模型科学能力的竞赛的基础上，其中包括新的人工智能科学奥林匹克竞赛，以吸引来自世界各地的杰出年轻人才进入该领域。除了直接影响之外，人工智能科学“希尔伯特问题”计划还可以为国际科学合作和资助提供一个受欢迎的焦点，并激励新一代跨学科科学家识别和解决人工智能问题。

2. Make the world readable to scientists

Most scientific data is uncollected, partial, uncurated or inaccessible, making it unavailable to train AI models. There is no single policy response to what is far from a uniform challenge. Policymakers and funders will need to blend a small number of top-down initiatives with support to scale up [promising grassroots efforts](#). A new international network of *AI for Science Data Observatories* should be set up to help address these goals. These Observatories could be given long-term backing and tasked with running rapid AI for Science ‘data stocktakes’, where expert teams map the state of data in priority disciplines and application areas. Stocktakes could identify existing datasets, such as the [Sequence Read Archive](#), whose quality could be further improved, as well as untapped or underutilised datasets, such as the decades of experimental fusion data that is currently unavailable to scientists or leading biodiversity datasets that are subject to restrictive licensing conditions. The stocktakes could also include new ‘data wish lists’. For example, our internal analysis suggests that less than 7% of papers in key environmental research domains use AI. We recently funded Climate Change AI to [identify datasets](#) which, if available or improved, could remove some of the bottlenecks to higher AI use. To ensure this analysis leads to action, policymakers should designate and empower organisations to be accountable for addressing the results of the data stocktakes.

The observatories could also scope the creation of new databases, including ensuring that adequate consideration is given to their long-term storage, maintenance, and incentives. This could include new databases to securely store the results of strategic wet lab experiments that are currently discarded, complemented by making the deposition of these experimental results a requirement for public research funding. Or digitising more public archives, following the example of a recent [UK government and Natural History Museum collaboration](#) to digitise their natural science collections, which includes more than 137 million items, from butterflies to legumes, across a 4.6 billion-year history. Policymakers can also empower scientists to use LLMs to create and improve their own datasets, by ensuring that publicly-funded research is open by default, where possible, building on recent examples from the [UK](#), [US](#) and [Japan](#), including mandates [to release research via pre-print servers](#). Policymakers could seek to co-fund the most ambitious dataset initiatives with industry and philanthropy. ■

3. Teach AI as the next scientific instrument

Over the past half century, as the number of scientific technologies has grown, so has most scientists’ distance from them. Many technologies are the products of science, but [an ever smaller share of scientists](#) are trained in how to develop and use them effectively. The pressing near-term need is to fund and incentivise mass uptake of shorter, more tactical AI training programmes [and fellowships](#), for existing scientists and research leaders. Policymakers can incentivise these efforts by setting a clear goal that every postgraduate science student should be able to access introductory courses on using AI in science, including on the most important tools in their domain, in the same way as basic statistics is often taught today. The type and depth of training needed will depend on an individual’s discipline and profile, and could range from basic introductory courses about how to reliably use LLMs for everyday research tasks, through to more advanced courses on how to fine-tune AI models on scientific data, as well as how to address more complex challenges, such as evaluating whether the data they used to test a model’s performance has intentionally or unintentionally ‘leaked’ into the data used to train it.

2. 让世界对科学家来说是可读的

大多数科学数据都是未收集的、部分的、未经整理的或无法访问的，因此无法用于训练人工智能模型。对于远非统一的挑战，没有单一的政策回应。政策制定者和资助者需要将少量自上而下的举措与支持相结合，以扩大有希望的基层努力。应建立一个新的科学数据观测站人工智能国际网络来帮助实现这些目标。这些观测站可以得到长期支持，并负责运行快速的人工智能科学“数据盘点”，其中专家团队绘制优先学科和应用领域的数据状态图。盘点可以识别现有数据集，例如质量可以进一步提高的序列读取存档，以及未开发或未充分利用的数据集，例如目前科学家无法获得的数十年的实验融合数据或受限制性许可条件。盘点还可能包括新的“数据愿望清单”。例如，我们的内部分析表明，关键环境研究领域的论文中只有不到 7% 使用人工智能。我们最近资助了气候变化人工智能来识别数据集，如果这些数据集可用或得到改进，可以消除更高程度人工智能使用的一些瓶颈。为了确保这种分析能够带来行动，政策制定者应该指定并授权组织负责处理数据盘点的结果。

观察站还可以确定新数据库创建的范围，包括确保充分考虑其长期存储、维护和激励。这可能包括新的数据库，用于安全地存储目前被丢弃的战略湿实验室实验的结果，并通过将这些实验结果的沉积作为公共研究资金的要求来补充。或者将更多的公共档案数字化，效仿最近英国政府和自然历史博物馆合作将其自然科学藏品数字化的例子，其中包括超过 1.37 亿件物品，从蝴蝶到豆类，跨越 46 亿年的历史。政策制定者还可以授权科学家使用LLMs创建和改进自己的数据集，方法是确保公共资助的研究默认开放，在可能的情况下，以英国、美国和日本最近的例子为基础，包括要求通过预印本服务器发布研究成果。政策制定者可以寻求与行业和慈善机构共同资助最雄心勃勃的数据集计划。

3. 将人工智能教授为下一个科学工具

在过去的半个世纪里，随着科学技术数量的增长，大多数科学家与它们的距离也越来越远。许多技术都是科学的产物，但接受过如何有效开发和使用这些技术的培训的科学家却越来越少。近期紧迫的需求是资助和激励现有科学家和研究领导者大规模接受更短、更具战术性的人工智能培训项目和奖学金。政策制定者可以通过设定明确的目标来激励这些努力，即每个理科研究生都应该能够学习在科学中使用人工智能的入门课程，包括其领域中最重要的工具，就像今天经常教授基础统计学一样。所需培训的类型和深度将取决于个人的学科和个人资料，范围可以从有关如何可靠地使用LLMs进行日常研究任务的基本入门课程，到有关如何精细化的更高级课程。根据科学数据调整人工智能模型，以及如何解决更复杂的挑战，例如评估用于测试模型性能的数据是否有意或无意地“泄漏”到用于训练模型的数据中。

These programmes could build on established examples such as the University of Cambridge's [Accelerate Programme](#) that provides structured training in AI to PhD and postdoctoral researchers, or the [short courses](#) that The Carpentries offer on the programming, data, and computational skills needed to do research.

Policymakers also need to move quickly to put in place longer-term programmes to ensure that the next generation of scientists has the skills they need. This means mainstreaming and deepening AI training and skills development in science education at all levels. Secondary school science students will need [early exposure](#) to the impact of AI while university students will need access to [new types of interdisciplinary AI science degrees](#), such as the [pan-African AI for Science Masters programme](#) that [we partnered with](#) the African Institute for Mathematical Sciences to develop. [Dedicated scholarships](#) could also help. For example, the UK's BIG Scholarships programme provides [outstanding opportunities](#) to high school students, with a focus on those from underrepresented groups who have excelled in International Science Olympiads and want to continue their study in leading science hubs but lack the funds to do so. ■

4. Build evidence and experiment with new ways of organising science

Scientists' use of AI is growing exponentially, but policymakers have little evidence about who is doing it best, how they are doing it, and the hurdles that are inhibiting others. This evidence gap is an impediment to identifying the best AI for Science policy ideas and targeting them effectively. Historically, answers to such questions often come from fields such as economics or innovation studies, but the results can [take years to arrive](#). We are using citation data analysis, interviews, and community engagement to understand how scientists are using our AI models. Governments are [also investing](#) in these *metascience* capabilities to improve how they fund, share and evaluate science research. Building on this momentum, scientists could be tasked with a mission to rapidly assess foundational policy questions, including: where is the most impactful AI for Science research occurring and what types of organisations, talent, datasets, and evaluations are enabling it? To what extent are scientists using and fine-tuning LLMs vs more specialised AI models, and how are they accessing these models? To what extent is AI *actually* benefiting or harming scientific creativity, reliability, the environment, or other domains? How is AI affecting a scientist's perception of their job and what skills, knowledge gaps, or other barriers are preventing their broader use of AI?

Beyond informing robust policy responses, this evidence base will arm policymakers with the foresight they need to anticipate how AI will transform science and society, similar to the foresight they are developing for AI safety risks through the growing network of [AI Safety Institutes](#). The evidence will also highlight opportunities to reimagine the incentives and institutions needed for science in the age of AI. In particular, scientists and policymakers have [only explored a small fraction](#) of the possible approaches to organising and executing science research. The rise of AI provides a welcome forcing function to experiment with new types of institutions, from those with more freedom to pursue [high-risk, high-reward](#) research, to Focused Research Organisations aimed at addressing specific bottlenecks. And from new types of [interdisciplinary AI science institutes](#) in priority domains such as climate or food security, to completely novel institutions that we are yet to imagine. Those who experiment faster will stand to benefit the most from a new golden age of discovery. ■

这些项目可以建立在已有的例子的基础上，例如剑桥大学的加速项目，该项目为博士和博士后研究人员提供人工智能方面的结构化培训，或者 The Carpentries 提供的关于研究所需的编程、数据和计算技能的短期课程。

政策制定者还需要迅速采取行动，制定长期计划，以确保下一代科学家拥有他们所需的技能。这意味着将人工智能培训和技能开发纳入各级科学教育的主流并深化。中学生理科学生需要尽早接触人工智能的影响，而大学生则需要获得新型跨学科人工智能科学学位，例如我们与非洲数学科学研究所合作的泛非洲人工智能科学硕士项目发展。专门的奖学金也能有所帮助。例如，英国的 BIG 奖学金计划为高中生提供了绝佳的机会，重点关注那些来自代表性不足群体的学生，他们在国际科学奥林匹克竞赛中表现出色，希望在领先的科学中心继续学习，但缺乏资金。

4. 建立证据并尝试组织科学的新方法

科学家对人工智能的使用呈指数级增长，但政策制定者几乎没有证据表明谁做得最好、他们如何做以及阻碍其他人的障碍。这种证据差距阻碍了确定最佳人工智能科学政策理念并有效瞄准它们。从历史上看，此类问题的答案通常来自经济学或创新研究等领域，但结果可能需要数年时间才能得出。我们正在利用引文数据分析、访谈和社区参与来了解科学家如何使用我们的人工智能模型。各国政府也在投资这些元科学能力，以改善资助、分享和评估科学的方式。在此势头的基础上，科学家可以承担快速评估基本政策问题的任务，包括：最有影响力的人工智能科学研究发生在哪，以及什么类型的组织、人才、数据集和评估正在支持它？与更专业的人工智能模型相比，科学家在多大程度上使用和微调LLMs，以及他们如何访问这些模型？人工智能实际上在多大程度上有益或损害了科学创造力、可靠性、环境或其他领域？人工智能如何影响科学家对其工作的看法？哪些技能、知识差距或其他障碍阻碍了他们更广泛地使用人工智能？

除了为强有力的政策反应提供信息外，这一证据基础还将为政策制定者提供预测人工智能将如何改变科学和社会所需的远见，类似于他们通过不断发展的人工智能安全研究所网络针对人工智能安全风险开发的远见。这些证据还将凸显重新构想人工智能时代科学所需的激励措施和机构的机会。特别是，科学家和政策制定者只探索了组织和执行科学的研究的可能方法的一小部分。人工智能的兴起为试验新型机构提供了令人欢迎的强制功能，从那些可以更自由地从事高风险、高回报研究的机构，到旨在解决特定瓶颈的重点研究组织。从气候或粮食安全等优先领域的新型跨学科人工智能科学研究所，到我们尚未想象的全新机构。那些

更快的实验将从新的发现黄金时代中获益最多。



