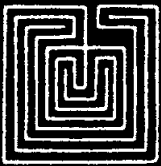


Dædalus

Journal of the American Academy of Arts & Sciences Winter/Spring 2026



AI & Science:
What Is the Future of Discovery?

James Manyika, guest editor

On the cover

Mapping Four Decades of AI Research: A UMAP Sculpture across Three *Dædalus* Issues

What does it look like when you compress nearly forty years of humanity's most rigorous thinking about artificial intelligence into a single form?

This work is a UMAP (Uniform Manifold Approximation and Projection) rendering of every word across three landmark issues of *Dædalus*—winter 1988, “Artificial Intelligence”; spring 2022, “AI & Society”; and winter/spring 2026, “AI & Science”—based on similarity clustering (six-dimensional: x, y, z and red, green, blue colorizing).

Three distinct moments when the Academy convened leading minds to take stock of where AI stands and where it's heading. The resulting image is not an illustration. It is a sculpture, shaped not by chisel or clay, but by the geometry of language itself, compressed through machine intelligence into visible form.

In 1988, the conversation was intimate and philosophical. Seymour Papert, Daniel Dennett, and Sherry Turkle debated whether machines could ever truly think. The vocabulary orbited the body: brains, neurons, perception, consciousness.

By 2022, the language had exploded outward. Fei-Fei Li, James Manyika, and Erik Brynjolfsson confronted what deep learning's success meant for justice, labor, and democracy.

The 2026 issue reaches further still. Demis Hassabis describes AI as an instrument for understanding reality itself—from protein folding to the laws of physics. Yann LeCun argues for world models that move beyond language into abstract representation.

The UMAP reveals the filaments between these eras. Words like *intelligence*, *learning*, and *human* appear in all three, but their neighbors shift—from *neuron* in 1988, to *equity* in 2022, to *prediction* in 2026. The dense clusters are shared concerns; the long tendrils reaching into darkness are ideas unique to each moment.

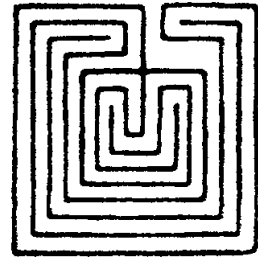
When Brunelleschi carved perspective into architecture, or when Michelangelo released figures from marble, they were not simply making objects—they were making visible the invisible structures of their age. The Renaissance sculptors gave form to human proportion, divine geometry, and anatomical truth. Today, a parallel undertaking is quietly emerging: the sculpting of data itself into physical and immersive form, using machine intelligence as both material and collaborator.

This is what I have tried to pursue in my own practice—humbly, and with deep respect for what came before—treating vast archives of information not as abstract numbers but as raw matter waiting to reveal its hidden shape. It is, I believe, the beginning of a new sculptural movement: one in which the artist's hand is guided by algorithms, in which the block of marble is replaced by multidimensional space, and in which what emerges is not a representation of the world but the actual structure of collective knowledge made tangible.

Three issues. Four decades. One continuous, expanding form.

— Refik Anadol
<https://refikanadol.com>

Dædalus



Journal of the American Academy of Arts & Sciences

“AI & Science: What Is the Future of Discovery?”

Volume 155, Number 1–2; Winter/Spring 2026

James Manyika, Guest Editor

Phyllis S. Bendell, Editor in Chief

Peter Walton, Senior Editor

Maya Robinson, Assistant Editor

Contents

- 7 **Introductory Notes: On AI, Science & the Future of Discovery**
James Manyika

In Dialogue: AI Pioneers on AI & Science

- 34 **AI as the Ultimate Tool for Science:
A Conversation with Demis Hassabis**
Demis Hassabis & James Manyika
- 45 **Learning Abstractions: A Conversation with Yann LeCun**
Yann LeCun & James Manyika

AI in the Life Sciences & Physical Sciences

- 61 **From Alchemy to Alchemy: On Matter, Minds & Tools**
Alán Aspuru-Guzik
- 78 **Unlocking Scientific Intuition & Reasoning at Digital Speed**
Pushmeet Kohli
- 92 **Beyond Representation: AI in Cellular Discovery**
Charlotte Bunne & Aviv Regev
- 110 **Building the Drug Discovery Engine of the Future
with AI-Empowered Nodal Biology**
Anna Greka
- 122 **From Pixels to Minds:
Mapping & Understanding the Brain with AI**
Viren Jain & Jeff Lichtman
- 136 **The Algorithmic Planet**
Anna M. Michalak & John C. Platt
- 151 **AI Reaches for the Stars**
Stella S. R. Offner

The Science of AI & Developing AI for Science

- 166 Language Is Not All You Need . . . but Language,
Probabilistic Programs & Bayesian Models of Cognition
Will Get You Pretty Far
Joshua B. Tenenbaum
- 185 Knowledge-Centric AI for Scientific Discovery
Carla P. Gomes
- 202 Building an AI Polymath
Shirley Ho
- 216 How Do We Build AI to Push the Frontiers of
Scientific Discovery?
Anima Anandkumar
- 229 Toward a Science of Intelligence:
Unifying Physics, Neuroscience & AI
Surya Ganguli
- 246 Quantum + AI = Quantum AI
Maria Spiropulu & Hartmut Neven

The Social Science of AI for Science

- 266 The Future of AI-Facilitated Medicine
Eric J. Topol
- 277 The Role of AI in Drug Discovery in Africa
Kelly Chibale
- 289 Physics Is Different: Context, Culture & Craft
in Effective AI for Physics
Tess Smidt

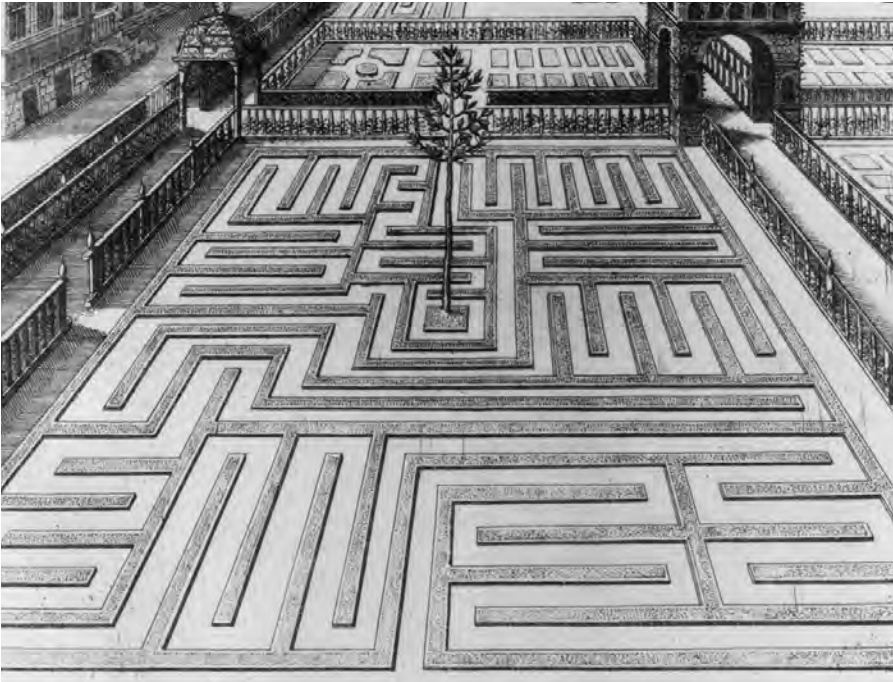
- 306 **Thinking & Doing Science in the Age of AI**
Alison Noble
- 316 **Field Theory: AI as Social Science Question, Object & Tool**
Alondra Nelson
- 335 **Philosophy of Autonomous Science: Ten Questions for
the Coming Age of Artificial Scientists**
Mario Krenn & Heather Champion

Personal Briefs

- 350 **Geometry-Informed AI for Scientific Discovery**
Melanie Weber
- 354 **Are Current AI Systems Unlocking Knowledge Discovery
in Genomics?**
Antonio Orvieto
- 358 **AI & the Discovery of Molecules through
Autonomous Laboratories**
Connor Coley
- 361 **Scaling Physics Intelligence for the Earth's Subsurface**
Gege Wen
- 364 **AI & Ecology: From Tool to Transformation**
Sara Beery
- 368 **Making Automation Work for Social Scientists**
M. J. Crockett
- 373 **When AI Meets Art at the Scales of Science**
Refik Anadol

Dædalus

Journal of the American Academy of Arts & Sciences



Design for the hedge labyrinth is by Johan Vredeman de Vries, from *Hortorum viridariorumque elegantes & multiplices formae: ad architectonicae artis normam affabre delineatae* (Cologne, 1615).

Dædalus was founded in 1955 and established as a quarterly in 1958. Its namesake was renowned in ancient Greece as an inventor, scientist, and unriddler of riddles. The journal's emblem, a labyrinth seen from above, symbolizes the aspiration of its founders to "lift each of us above his cell in the labyrinth of learning in order that he may see the entire structure as if from above, where each separate part loses its comfortable separateness."

The American Academy of Arts & Sciences, like its journal, brings together distinguished individuals from every field of human endeavor. It was chartered in 1780 as a forum "to cultivate every art and science which may tend to advance the interest, honour, dignity, and happiness of a free, independent, and virtuous people." Now in its third century, the Academy, with its more than five thousand members, continues to provide intellectual leadership to meet the critical challenges facing our world.

Dædalus Winter/Spring 2026
Issued as Volume 155, Number 1–2

© 2026 by the American Academy of Arts & Sciences.

Cover artwork © 2026 by Refik Anadol

Mapping Four Decades of AI Research © 2026 by Refik Anadol

Machine Memoirs: Space © 2021 by Refik Anadol Studio

Sense of Space: Molecular Architecture © 2021 by Refik Anadol Studio

Sense of Space: Connectome Architecture © 2021 by Refik Anadol Studio

Unsupervised—Machine Hallucinations © 2022 by Refik Anadol Studio

Editorial offices: *Dædalus*, American Academy of Arts & Sciences, 136 Irving Street,
Cambridge MA 02138. Phone: 617 576 5085. Email: daedalus@amacad.org.

Library of Congress Catalog No. 12-30299.

Dædalus publishes by invitation only and assumes no responsibility for unsolicited manuscripts. The views expressed are those of the author(s) of each essay, and not necessarily of the American Academy of Arts & Sciences.

Dædalus (ISSN 0011-5266; E-ISSN 1548-6192) is published quarterly (winter, spring, summer, fall) by The MIT Press, 255 Main Street, Suite 9, Cambridge MA 02142, for the American Academy of Arts & Sciences. An electronic full-text version of *Dædalus* is available from amacad.org/daedalus and from The MIT Press.

Dædalus is published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license. For allowed uses under this license, please visit <https://creativecommons.org/licenses/by-nc/4.0>.

Members of the American Academy please direct all questions and claims to daedalus@amacad.org.

Advertising inquiries may be addressed to Marketing Department, MIT Press Journals, 255 Main Street, Suite 9, Cambridge MA 02142. Phone: 617 253 2866. Fax: 617 253 1709. Email: journals-info@mit.edu.

To request permission to photocopy or reproduce content from *Dædalus*, please contact the Subsidiary Rights Manager at MIT Press Journals, 255 Main Street, Suite 9, Cambridge MA 02142. Fax: 617 253 1709. Email: journals-rights@mit.edu.

Corporations and academic institutions with valid photocopying and/or digital licenses with the Copyright Clearance Center (CCC) may reproduce content from *Dædalus* under the terms of their license. Please go to www.copyright.com; CCC, 222 Rosewood Drive, Danvers MA 01923.

Printed in the United States by The Sheridan Press, 450 Fame Avenue, Hanover PA 17331.

Postmaster: Send address changes to *Dædalus*, 255 Main Street, Suite 9, Cambridge MA 02142. Periodicals postage paid at Boston MA and at additional mailing offices.

The typeface is Cycles, designed by Sumner Stone at the Stone Type Foundry of Guinda CA. Each size of Cycles has been separately designed in the tradition of metal types.

Introductory Notes: On AI, Science & the Future of Discovery

James Manyika

The last few years have seen a flurry of major scientific awards at the intersection of AI and science: In 2024, the Nobel Prize in Physics was awarded to Geoffrey Hinton and John Hopfield for their pioneering work on machine learning with artificial neural networks, which has formed the basis of modern AI, while the Nobel Prize in Chemistry recognized David Baker, Demis Hassabis, and John Jumper for their work on protein design and structure prediction. In 2025, the Nobel Prize in Physics went to John Clarke, Michel Devoret, and John Martinis for laying the foundations of quantum computing and quantum AI. The Turing Award (“the Nobel for Computer Science”) was conferred in 2024 on Andrew Barto and Richard Sutton for their contributions to reinforcement learning and, not long before that, in 2018 on Yoshua Bengio, Geoffrey Hinton, and Yann LeCun for their pioneering work on deep learning. In addition to the achievements, these awards – spanning multiple scientific disciplines and drawing on ideas from the different disciplinary backgrounds of the recipients – underscore the bidirectional nature of progress at the intersection of AI and science: science advancing AI and AI advancing science.

The work by Baker and by Hassabis and Jumper is notable not only for *what* was achieved but also *how* it was achieved – and for what *that* portends for the future of discovery.¹

With AlphaFold, Hassabis and Jumper solved the fifty-year grand challenge of protein-structure prediction. In his 1972 Nobel lecture, biochemist Christian Anfinsen had postulated that a protein’s structure is fully determined by its amino acid sequence. But predicting this shape computationally was nearly impossible because a typical protein has an astronomical number of possible configurations, far beyond what could simply be searched by brute force. By achieving experimental accuracy, AlphaFold demonstrated that AI could learn the “grammar” of biology from data, bypassing the need for computationally expensive *ab initio* simulations. The team released the predicted structures of over two hundred million proteins – nearly the entire known protein universe – which have now been used

by more than 3.3 million scientists in over 190 countries. Baker's work asked the inverse question: Can we design a protein that binds to a specific target? He and his team adapted AI diffusion models to start with random amino acid coordinates and then "denoise" them to form a valid protein backbone that fits a specific constraint. For example, if a researcher wants a protein that binds to a specific site on a virus, they condition the diffusion model on that site and the model "hallucinates" (a useful phenomenon in this case) a protein shape that perfectly complements it.² In experimental validation, the success rate of proteins that bind skyrocketed by orders of magnitude compared with previous methods, making possible the design of proteins that were formerly impossible to engineer.

To these AI-enabled advances we could add many more from just the last few years in genomics, neuroscience, health care, chemistry, materials science, plasma physics, astrophysics, climate science, and beyond.³ To be clear, the now constant stream of announced advances has not been without instances of hyperbole, overstatement, false claims, *in silico* (computer simulated) breakthroughs that don't match real-world reality, contested results, and outright failures. In this regard, AI is like every paradigm change that preceded it.

The combination of progress in AI and its related techniques, its expanding and general usefulness in science, and its celebrated landmark contributions to discovery has led in recent years to recognition by the scientific community that a new era of AI for science has arrived. Indeed, many universities, research consortia, national academies, national laboratories, and governments have launched AI-for-science initiatives.⁴ In the 2023 *Nature* article "Scientific Discovery in the Age of Artificial Intelligence," a group of prominent scientists put the matter succinctly: AI has now transcended its role as a peripheral tool for analysis to become a foundational mechanism for discovery.⁵

Yet there are also challenges and an unease about the shifts accompanying this new era: for example, that academic and national laboratories are ill-suited for the "data-intensive and compute-intensive training regime of AI models" and face a brain drain to the private sector, where most frontier AI research and AI-enabled science now takes place (though often involving academic collaborators). Many have raised concerns about AI's impact on *how* science is done, from reproducibility and contamination of the scientific literature to epistemic challenges and expanded safety and ethical considerations.⁶ Nevertheless, most scientists acknowledge that AI-enabled science is here to stay and that AI should be harnessed to advance science, perhaps with a broader or reimagined view of *what* could be achieved and *how*.

So, what is the future of scientific discovery in the age of AI?

The most recent volume of *Dædalus* that focused on artificial intelligence was published in spring 2022. That volume, which I had the honor of editing, focused

on AI & Society. It was not the first *Dædalus* issue to focus on AI – that distinction goes to a volume published in 1988. The 1988 *Dædalus* volume captures the tension at the time: on one hand, disappointment that AI had not lived up to the high expectations set at the field’s founding in the 1950s (typified by Hilary Putman’s essay “Much Ado About Not Very Much”); on the other, a sense that the field was on the verge of a rebirth. The 2022 volume explores the renewed technical progress in AI as well as its implications for society, from jobs and the economy to democracy and distrust, governance, national security, justice, and ethics.

Much has happened since the publication of that 2022 volume of *Dædalus*: It came out six months before the launch of ChatGPT demonstrated to the public just how far AI had come. Progress in AI has continued at a rapid pace since, as chronicled in numerous reports, such as our annual Stanford AI Index reports in this period (2022–2026).⁷ With this progress, its intersections with society have expanded, resulting in more excitement about AI’s useful and beneficial possibilities as well as increased concern about the challenges and risks. The stakes – positive and negative – have grown in both directions: from AI’s impact on what individuals are now able to do, to its effects on productivity and growth in the economy, and on jobs, education, and creativity; to its use, misuse, missed use, and the societal implications; to its infrastructure and energy footprint, its geopolitics, its regulation.⁸ These developments warrant deeper assessment and discussion. But this is not that volume; this volume is focused on AI and science.

For science, this moment is not unlike that before the arrival of ChatGPT in 2022. Those at the dual frontiers of AI and its use in science recognize how AI is advancing science, but the general public remains much less aware. This gap stems partly from the indirect nature of the impact on individuals and society of AI’s contributions to science – unlike the firsthand experiences that many people have with chatbots like Claude, ChatGPT, or Gemini – and also from the topicality and, some would argue, immediacy of AI’s other aforementioned impacts. When it comes up at all, the prospect of AI advancing science is often touted as something to look forward to, a reason to take AI seriously, the pot of gold at the end of the rainbow. That view is correct in that many more of the possibilities of AI for science lie in the future, but it misses its ongoing contributions to science and the advances *already* being achieved, as well as the extent to which society is already benefiting in areas ranging from health care to weather forecasting to food security.

While this *Dædalus* volume may help address gaps in awareness, that is not its primary aim. Rather, its main motivation is to explore the future of scientific discovery in the age of AI. Given the unwieldy scope this suggests, this collection focuses on and draws primarily from scientists working at the frontier of the intersection of AI and science, while acknowledging that this does not represent the full range of activities or views in the scientific community (nor does it cover all the issues that our core question conjures).

The volume consists of contributions from thirty-three scientists, spanning the life sciences and medicine, cognitive science and neuroscience, the physical and earth sciences, chemistry and materials science, computer science, mathematics, and the social sciences. Their contributions are organized into five sections: The first engages two notable scientists at the frontier of AI's advance. The second section, "AI in the Life Sciences & Physical Sciences," explores what's being achieved and what lies ahead. The third, "The Science of AI & Developing AI for Science," examines AI's current limitations and efforts to advance it to further enable discovery. The fourth, "The Social Science of AI for Science," explores the wider implications of AI-assisted science, including on how science is done, the role of the scientist, the scientific method, and understanding. The final section consists of "personal briefs" from early-career scientists for whom the question "Are you an AI researcher or a scientist?" feels outdated, as well as a statement by an artist on "When AI Meets Art at the Scales of Science." Despite these groupings, each contribution should be read as a unique response to this volume's motivating question. I am grateful to the contributors for accepting my invitation to participate in this volume.

As when I edited the 2022 *Dædalus* issue on AI & Society, I am not a bystander – not in AI generally nor its intersection with science specifically. I lead research teams in these areas, have coauthored related papers (some cited in this volume), and am involved in relevant work in academia. Indeed, my research in AI started in the 1990s and focused on solving scientific questions. Undoubtedly, this volume is colored not only by my various involvements but, most importantly, by the views of its contributors. To supplement these perspectives, I have included extensive endnotes to facilitate further exploration.

Before proceeding, it may be useful to define AI. As noted in my introduction to the 2022 *Dædalus*, the field has a historical grounding in the now-famous 1956 Dartmouth Summer Research Project on Artificial Intelligence, to which the term "artificial intelligence" is largely attributed.⁹ In that volume, I also noted what Marvin Minsky had earlier pointed out: that AI had become a "suitcase word" packed variously with a wide range of concepts and techniques.¹⁰ That problem has since worsened. In our context of AI and science, the term encompasses a Minsky-ean suitcase of approaches, from statistical, computational, and machine-learning and deep-learning techniques to analyze, infer, and discern patterns, learn from data, form abstractions of sensory inputs, build models of natural phenomena, and make predictions, to approaches and systems that simulate physical laws, explore possibility spaces, generate hypotheses, propose designs, and steer experiments, including through the human-like capabilities (such as reasoning) that those at Dartmouth had in mind.

It is sometimes assumed that the intersection of AI and science is something that took-off with the arrival of large language models (LLMs). Therefore, a brief history that selectively signposts the lineage of a few ideas that underpin and continue to animate this intersection may be useful.

For centuries, scientific discovery rested on two pillars: *theoretical science*, rooted in the deductive logic of mathematics and natural philosophy, seeking to construct models of reality – from Newton’s laws to Einstein’s equations – with predictive power; and *empirical science*, grounded in observation and experimentation serving as the arbiter of a theory’s validity. The latter half of the twentieth century saw the crystallization of a third pillar: *computational science*.¹¹ This did not merely automate arithmetic and extend the power of the former two; it brought forward a fundamentally distinct methodology for investigating phenomena that were analytically intractable, experimentally inaccessible, or both.¹² While this may have crystallized midcentury with the advent of digital computers, the history of computational science stretches back to antiquity (with much of the computation painstakingly human, occasionally mechanical).¹³ That history is bidirectional: computational approaches advancing science and science advancing computational approaches, with pioneering scientists from multiple disciplines operating on this two-way street right up to the present with AI.

The polymaths Alan Turing and John von Neumann stand out in this lineage of ideas.¹⁴ Among its various insights, Turing’s seminal 1936 paper “On Computable Numbers” suggests that if reality is based on scientific laws and those laws are computable, then reality could be modeled as a Turing machine and the nature of reality could be informational – an idea espoused by Demis Hassabis and some others.¹⁵ Turing later proposed that a network of randomly connected nodes (like those in the brain) could be trained rather than programmed and could organize itself for a definite purpose. He then inverted the problem: Instead of using biology to inspire AI, how could the computer be used to elucidate a biological question? He showed that biological patterns could be explained by the interaction of simple chemical substances, which he modeled using a reaction-diffusion model. Taking a different tack, in his 1958 paper “The Computer and the Brain,” von Neumann compares the artificial and the natural automaton, concluding that the computer (artificial) was high-speed and high-precision, while the brain (natural) was a slow, low-precision mixed system that was probabilistic and massively parallel. Motivated by the question of how complex organisms evolve from simpler ones without violating the second law of thermodynamics, he introduced in 1966 his “Theory of Self-Reproducing Automata.” In the 1980s, computer scientist and physicist Stephen Wolfram would develop this further and show that simple automata offered a computational alternative to the differential equations that had dominated physics.¹⁶

In the wake of Turing and von Neumann, the pursuit of AI branched in the 1950s into two competing visions – the “symbolic” (mind) approach and the “connectionist” (brain) approach – with science a primary proving ground. The symbolic approach, championed by Allen Newell, Herbert Simon, and Marvin Minsky, defined intelligence as symbol manipulation according to formal rules and heuristics. In 1956, their Logic Theorist algorithm demonstrated that a machine could

prove mathematical theorems – specifically, thirty-eight theorems from *Principia Mathematica* – using heuristic search rather than brute force.¹⁷ This suggested that if scientific knowledge could be axiomatized, machines could theoretically derive all consequences of those axioms and essentially compute the truths of the universe. On the other branch, in 1951, Marvin Minsky and Dean Edmunds built SNARC, the first connectionist neural network to demonstrate reinforcement learning; it simulated a rat navigating a maze. And in 1957, Frank Rosenblatt developed the perceptron, a probabilistic model capable of learning from data. However, such learning models were overshadowed by the success of general problem solvers built on heuristics, logic, and knowledge of science.

By the 1970s, as it became apparent that general problem solvers could not scale to complex scientific domains in which the combinatorial explosion overwhelmed simple logic, AI had already begun to shift to domain-specific approaches. For example, Margaret Dayhoff's key insight that biology was inherently an information science allowed her to pioneer probabilistic models for comparing protein sequences, enabling the construction of phylogenetic trees and the study of molecular evolution. DENDRAL, started in 1965 by Edward Feigenbaum, Joshua Lederberg, and Bruce Buchanan, used heuristics derived from chemists to analyze mass spectrometry data and elucidate molecular structures. Its successor, Meta-DENDRAL, became the first to automate discovery by inferring new cleavage rules for mass spectrometry from data. But by the late 1980s, such expert systems were hitting a wall. Extracting rules from experts was slow and expensive, and the systems couldn't handle edge cases. Simultaneously, scientific datasets were growing too large for manual rule creation but were perfect for probabilistic analysis.¹⁸

Meanwhile, the field was rediscovering the learning approach to AI. In 1982, John Hopfield developed a model of associative memory by drawing an analogy to magnetic spin systems in condensed matter physics and statistical mechanics. The Hopfield model explained how networks of neurons could store and recall facts and led to the creation of Boltzmann machines, thereby connecting statistical mechanics with machine learning. This influenced Geoffrey Hinton's study of Boltzmann machines; thus, the progression from Hopfield models to Boltzmann machines and belief networks became the "arc of discovery" to deep learning, as Alán Aspuru-Guzik and Surya Ganguli discuss in this volume. Around the same time, while exploring how neural processing works in the brain, David Rumelhart, Geoffrey Hinton, and Ronald Williams were using back-propagation to train neural networks. They cited Yann LeCun, who by 1989 was applying back-propagation to handwritten zip code recognition, a precursor to machine vision. It was also during this period that Andrew Barto and Richard Sutton laid the foundations of modern reinforcement learning, drawing from fields as diverse as psychology to control theory.¹⁹ This period represented the rebirth signaled by the 1988 *Dædalus* volume, especially in the essay by Hubert Dreyfus and Stuart Dreyfus,

“Making a Mind Versus Modeling the Brain : Artificial Intelligence Back at a Branch-point.”²⁰

In the 2022 *Dædalus* volume on AI & Society, Jeffrey Dean describes the period since 2012 as a “golden decade” for AI, driven by advances in deep learning, computing (chips, infrastructure, and tools) suited to deep learning, and the explosion of data to train models.²¹ In that same volume, Kevin Scott highlights work decades earlier – by Ray Solomonoff and Claude Shannon linking entropy, information, and probability theory with machine learning – that now enables deep-learning models to compress and generalize what they learn.²² In this volume, LeCun adds additional insights on AI’s progress during this period, while Aspuru-Guzik as well as Charlotte Bunne and Aviv Regev note AI’s contributions to biology and chemistry.²³ AI for science in this era came to mean the use of deep neural networks as universal function approximators to learn from data and model complex scientific phenomena, make predictions, and accelerate simulation. The Nobel-worthy breakthroughs that I highlighted at the beginning became defining scientific achievements of this era.

So where are we now, and what is the future of discovery in the age of AI?

What is the perspective from the AI frontier?

Outside the field, it is sometimes assumed that those at the frontier of developing AI agree on everything, or that there is consensus on its technical and scientific trajectories. The dialogues with Nobel laureate Demis Hassabis and Turing awardee Yann LeCun that open this volume illustrate that this is not the case. In these conversations we find agreements, yes, but also differences and debates on the limits of LLMs, scaling, and tokenization, further breakthroughs needed in AI, artificial general intelligence (AGI), how AI can advance science and vice versa, and the future of discovery and open science.

While readers can engage directly with the perspectives of Hassabis, LeCun, and other contributors, it may also be useful to sample the views of other pioneering scientists to provide context for this volume and as inducement to explore further.

Dario Amodei, a biophysicist and cofounder of Anthropic, has stated that the “scaling hypothesis” is an empirical fact, that current transformer architectures, scaled sufficiently, will eventually master causal mechanisms of, for example, biology simply by observing enough biological data. Amodei is concerned, however, with interpretability and sees it as critical for scientific understanding.²⁴ As counterpoint, Ilya Sutskever, a pioneering researcher and cofounder of OpenAI and Safe SuperIntelligence, once an advocate of scaling, recently argued that the “age of scaling” is hitting diminishing returns and a new paradigm is needed, observing that “we are back in the age of wonder and discovery.” Like some others in this volume, he advocates for System 2 approaches as critical for science, which requires deep, multistep reasoning rather than just next-token prediction.²⁵ Fei-Fei Li, another

pioneering computer scientist in AI, believes what's required are world models to reason about three-dimensional geometry and physics, as well as a body to interact with the physical world.²⁶

Turing awardee Yoshua Bengio stands in contrast to the growing agentic vision for AI. He argues that until we solve alignment – ensuring that AI systems' goals, behaviors, and actions match human preferences and goals – giving AI agency (the ability to pursue goals in the world) is risky, even in a scientific context. He is developing “Scientist AI,” a system designed to understand the world, generate explanatory theories, and answer questions, thus restricting AI to epistemic rather than instrumental tasks.²⁷ Stuart Russell, computer scientist and coauthor (with Peter Norvig) of the most read textbook on AI, also argues for non-agentic or “oracle” architectures, citing alignment and safety risks. Russell believes that AI can take science beyond “Edisonian” trial-and-error and move fields like biology from a descriptive science to a predictive one. Describing LLMs as “circuitry” without “concepts,” he sees gaps that are critical for science.²⁸

In perhaps the most provocative and historically grounded thesis, Richard Sutton, another Turing awardee, has offered what he calls “the bitter lesson”: that every time researchers tried to build human knowledge into AI, it worked in the short term but blocked long-term progress; and that by encoding our current scientific theories into AI, we bias it toward what we already know and thus prevent it from discovering radically new physics or biology. Sutton argues we must instead build general learning methods, including AI systems capable of search and “meta-methods” to discover science on their own. He states that LLMs on their own are a dead end because they are static, that true scientific AI must learn and interact with the world to discover new laws.²⁹ It's worth noting that one of the possibilities of the learning approach is self-improving AI. On one hand, self-improvement could lead to AI that is even more capable of advancing discovery; on the other hand, if self-improvement is recursive and outruns our capacity to ensure safety, alignment, and governance, it could create unprecedented risks. While the idea of recursive self-improvement is not new, the recent progress in AI has attracted some to pursue its development, and has sparked debates about its feasibility and concerns about its safety.³⁰

Such varied views suggest that while AI advances have led to the development of much that is useful, including for science, the question *What is the future of discovery?* applies equally to artificial intelligence as a scientific field itself.

How is AI advancing discovery, and what are the possibilities for the future?

In his contribution “From Alchemy to AIchemy: On Matter, Minds & Tools,” Alán Aspuru-Guzik explores the transition from empirical alchemy to twentieth-century integration of computational science with chemistry and materials science to the

AI-enabled advances of today and the future.³¹ He also describes the evolution of modalities of discovery with AI from tools to agents, such as *El Agente* for quantum chemistry, to self-driving laboratories with paradigm-shifting implications for scientific understanding. Using examples of advances in biology, materials science, plasma physics, and mathematics, in “Unlocking Scientific Intuition & Reasoning at Digital Speed,” Pushmeet Kohli frames AI not merely as a computational accelerator but as a novel epistemic instrument that will take discovery further to predictive design, generating hypotheses and orchestrating experiments. This would precipitate a profound shift in the role of the scientist: from solver of puzzles to “architect of questions.” Coupling his optimism with caution, Kohli invokes E. M. Forster’s *The Machine Stops* to warn against a passive reliance on systems we cannot interpret, arguing that while AI may collapse the distance between question and answer, humanity must maintain the *why* behind the question.

With an eye on breakthroughs in diseases, Bunne and Regev, in “Beyond Representation: AI in Cellular Discovery,” and Anna Greka, in “Building the Drug Discovery Engine of the Future with AI-Empowered Nodal Biology,” focus on cellular biology. They articulate distinct visions in which the “cell-prediction problem” becomes the fulcrum of discovery. Bunne and Regev concentrate on the emergence of “virtual cells”: computational surrogates built on AI models that integrate multi-omic and spatial modalities to simulate and explore cellular dynamics. Greka argues that while the field celebrates AI’s triumphs in drug design and clinical optimization, the rate-limiting step remains target identification. She is pursuing the cell-prediction problem by forecasting responses of any human cell to perturbations using AI models trained on massive cell datasets. Such “nodal biology” could advance the identification of mechanisms across many genetic pathologies. In “From Pixels to Minds: Mapping & Understanding the Brain with AI,” Viren Jain and Jeff Lichtman turn AI’s microscope to the brain’s architecture and the intractable complexity of mapping synaptic connectivity, where manual analysis of petascale electron microscopy data is impossible. By evolving from static anatomical catalogs to AI models that simulate neural dynamics and predict functional perturbations, they are charting a path to decoding the brain’s neural syntax of healthy circuits to develop fingerprints of pathologies such as schizophrenia and autism. In an inversion reminiscent of Turing, they are also using the AI-enabled nanoscale connectomics to investigate whether the brain’s neural circuits implement computational primitives analogous to those in AI systems, such as the attention mechanisms in transformers.

In “The Algorithmic Planet,” Anna Michalak and John Platt reframe climate action and understanding of the physical Earth as problems of constrained optimization. They distinguish between “tractable” problems and the “intractable frontier” of long-term projection in which nonstationarity and chaotic variability confound purely data-driven approaches. The tractable frontier is where AI is increasingly being deployed around the world, for example, optimizing power grids, avoiding

contrails, and improving predictions of extreme weather events like floods, hurricanes, and wildfires.³² With progress in Earth foundation models, Michalak and Platt weigh this revolution against the paradox of AI's growing carbon footprint – not unlike Gege Wen's "double Ouroboros" and Aspuru-Guzik's warning of a possible "AIocene."³³ Wen's double Ouroboros also suggests that to scale AI, we must reinvent energy, and to reinvent energy, we must use AI, as in her work navigating the intricate physics of the subsurface. Meanwhile in "AI Reaches for the Stars," Stella Offner details how AI has graduated from classifying celestial morphologies to accelerating computationally prohibitive simulations of galaxy evolution and beyond. Shirley Ho, in "Building an AI Polymath," and Mario Krenn and Heather Champion, in "Philosophy of Autonomous Science: Ten Questions for the Coming Age of Artificial Scientists," survey other notable advances in astrophysics. However, as Offner notes, there are profound epistemic challenges: for example, unlike in the terrestrial sciences, astronomical properties – mass, temperature, density – cannot be measured directly but must be inferred from observed light. This exacerbates the tension between scientific rigor and the "black box" opacity and predictive power of deep learning that many in this volume note. At the same time, AI may offer new avenues to tackle what some have characterized, for example, as the "crisis in particle physics," in which experimental results have failed to yield new insights for questions in the standard model. So rather than analyze data from new instruments to test and confirm existing theories, why not use AI to explore unencumbered by current theories?³⁴

What's clear is that many AI-enabled advances so far concern high-dimensional problems for which data are abundant, as opposed to areas governed by chaotic, nonstationary dynamics or those with sparse data, which for now remain the "intractable frontier."³⁵ Further, many examples illustrate what may be the most distinctive aspect of the AI-for-science era: the possibility of discovery at unprecedented scale, scope, and speed, such as taking on *all* proteins and cells, exploring vast combinatorial spaces (computationally out of reach with brute force methods), learning from multidimensional and multimodal data – even multidisciplinary theories and ideas – simulating dynamics and interactions, and probing and learning from perturbations *in silico*. Yet sometimes there are significant discrepancies between simulation or generated artifacts and the real world; as Connor Coley puts it, "every [*in silico*] proposal eventually collides with the physical world." Such discrepancies, together with the inconsistency or "jaggedness" of current AI, reveal the paradox that along with novel insights there is the possibility of artifacts not grounded in reality or hallucinations that could mislead or indeed inspire novel directions (such as in Baker's Nobel-worthy work). But as Turing observed, "If a machine is expected to be infallible, it cannot also be intelligent," suggesting that for an AI (or human mind) to be capable of novel discovery, it must possess the capacity to make inductive leaps, guesses, and mistakes, and learn from them. Thus,

in the hands of a discriminating scientist or one able to validate *in silico* results in a “wet lab,” this jagged AI can be usefully employed to advance science.³⁶

The formal sciences – distinguished by their abstract structures and axiomatic and logic formalism rather than empirical observation of the physical world – have also been a proving ground for AI. In computer science, there have been advances on some long-standing problems: for example, in matrix multiplication, sorting and hashing algorithms, combinatorics and computational complexity, and theorem proving.³⁷ What about mathematics? In their recent paper “Mathematical Exploration and Discovery at Scale,” Fields medalist Terence Tao and his colleagues find that while current systems like AlphaEvolve (which achieved gold medal–level performance at the International Mathematical Olympiad, as Kohli discusses) excel at “constructive mathematics at scale,” they encounter epistemological boundaries and falter when problems resist formulation as smooth, optimizable functions amenable to hill-climbing.³⁸ This is not unlike Melanie Weber’s view in her essay that AI is particularly effective for “needle in a haystack” problems in mathematics. Thus, while AI can discover complex “constructions,” it currently lacks the capacity for deep new insights, serving for now as a powerful prosthetic for human intuition rather than as an independent generator of novel theory.

Of the many examples of AI-enabled advances in science, one could ask *what* exactly is being achieved: Is it better and more-useful models or predictors of natural phenomena than was possible otherwise? Is it the scale, scope, and speed of AI, along with the ability to learn from and explore vast spaces, without which the advances would have been practically infeasible or taken unfathomably long? Is it truly novel discoveries or breakthroughs on “grand challenge” problems? Or is it also the new scientific understanding we gain as a result? The answer clearly varies across the examples highlighted in this volume and beyond, with the majority (for now?) exhibiting the first couple of characteristics. Perhaps more examples of the latter kind will emerge, as more scientists use AI on more long-standing or new problems and in new ways (more on this later), or as AI’s capabilities advance further.

How can science advance AI to further advance science?

In “Language Is Not All You Need,” Joshua Tenenbaum critiques the orthodoxy that general intelligence emerges solely from scaling LLMs. He contrasts the “jaggedness” of current AI – which achieves superhuman results in some complex problems yet fails at elementary reasoning – with the robust, general intelligence of biological organisms. Positing that cognition evolutionarily precedes language, Tenenbaum argues that the field must transcend “scaling up” and emulate the biology of “growing up” (with echoes of Turing and Solomonoff). His synthesis of LLMs with Bayesian cognitive science and probabilistic programming aims to achieve the causal inference and generalization necessary for science. Going in the

other direction, he argues that such AI approaches may also “start to answer some of the fundamental questions about how our minds work and where our intelligence comes from.” On the limits of language, it is worth noting Antonio Orvieto’s critique in “Are Current AI Systems Unlocking Knowledge Discovery in Genomics?” that LLMs struggle with “state tracking” over the extremely long sequences found in genomics. Orvieto advocates for abandoning text-centric paradigms in favor of neural networks designed to achieve Turing-complete reasoning on biological data.³⁹ In her essay “Knowledge-Centric AI for Scientific Discovery,” Carla Gomes contends that the prevailing data-centric paradigm of LLMs struggles to generalize beyond training distributions and adhere to inviolable scientific principles. By embedding scientific constraints directly into deep-learning architectures, she highlights novel advances in fields like materials science and ecology, including some that elude human intuition.

Progressing toward “Building an AI Polymath,” Shirley Ho tackles a different challenge in which specialized AI models excel in specific domains yet fail to grasp nature’s interconnectedness. Invoking Da Vinci and Newton, she is developing foundation models capable of reasoning across domains and scales, from quantum mechanics to fluid dynamics and astrophysics (see also Stella Offner’s, Gege Wen’s, and Sara Beery’s perspectives on scales in their essays). Ho uses universal tokenization to put heterogeneous data archetypes into a unified latent space and uses symbolic distillation to translate neural computations into human-readable physical laws. Anima Anandkumar redirects our gaze toward the continuous nature of physical reality in her essay “How Do We Build AI to Push the Frontiers of Scientific Discovery?” She uses neural operators to go beyond discretization to learn mappings between continuous function spaces, allowing AI to internalize physical laws at multiple scales. Her stance offers a counterpoint to universal tokenization, asserting that the resolution of scientific discovery must be continuous and infinite.

In Surya Ganguli’s “Toward a Science of Intelligence: Unifying Physics, Neuroscience & AI” and Maria Spiropulu and Hartmut Neven’s “Quantum + AI = Quantum AI,” the authors contend that further progress in AI requires a reexamination of its physical and biological substrates. Like Tenenbaum, Ganguli heeds the “taunting sirens” of biology, that human brains outperform AI by orders of magnitude in energy and data efficiency. He observes that unlike digital systems that require energy-intensive, reliable bit-flips, biological computation thrives on slow, unreliable, stochastic, and even sloppy intermediate steps, and suggests that this may constitute a parsimonious design feature that could benefit AI development, rather than a bug. And in a synthesis of neuroscience and condensed matter physics, he envisions neuromorphic architectures that aid discovery broadly and may even help elucidate consciousness. Spiropulu and Neven look to the fundamental nature of reality – building on foundations laid by the winners of the 2025 Nobel Prize in Physics and recent progress in quantum computing, in part enabled

by AI – and posit that the trajectory of AI inexorably leads to “Quantum AI.” For example, they leverage quantum entanglement to develop out-of-time-order correlators (OTOCs) for elucidating molecular dynamics with novel scientific results.⁴⁰ Together, these essays suggest that for AI to be more capable for discovery, we will need architectures that reflect the fundamental nature of reality itself.⁴¹

Beyond this volume, many researchers are continuing to investigate the foundations of deep learning and its architectures, while others are building general assistive tools for scientists, many employing *in silico* agentic approaches, including literature-spanning multi-agent brainstorming, evolutionary mutations and tournaments, harnesses of various kinds, and neurosymbolic and surrogate modeling hybrids.⁴² Yet others are focused on more foundational barriers to scientific discovery. Among the most formidable barriers is what Judea Pearl called the “ladder of causation.”⁴³ Current deep learning is mostly on the first rung: *association*. Scientific discovery requires climbing to the rungs of *intervention* and *counterfactuals* (or imagining). One notable approach is Bernhard Schölkopf and his colleagues’ causal representation learning, which merges the feature-learning of deep neural networks with the rigor of causal inference, an approach being applied, for example, to disentangle the genetic drivers of disease from environmental confounders.⁴⁴ But perhaps the hardest aspect of doing science (and for current AI systems) is coming up with novel conjectures and theories. Physicist Max Tegmark has argued that a model that predicts planetary orbits with 99.9 percent accuracy but cannot derive Kepler’s laws is a tool for engineering, not physics. He and colleagues have been developing AI Feynman, a recursive symbolic regression algorithm to rediscover physical laws from data. In a long-standing tradition of testing AI via rediscovery of known science, in 2020, AI Feynman rediscovered one hundred equations from the Feynman Lectures on Physics.⁴⁵ Such directions suggest the possibility of novel theories and conjectures in a human-interpretable form.

What are the implications of AI-enabled science?

The integration of AI into the scientific enterprise represents more than enablement of discovery; it is also arguably a fundamental reconstitution of the social science of discovery – its applications and impacts on society, its institutions and people, its epistemic foundations and methods.⁴⁶ Most of the contributions to this volume engage these issues to varying degrees, some directly.

Eric Topol and Kelly Chibale focus on health care. In his essay “The Future of AI-Facilitated Medicine,” Topol envisions a medical renaissance in which AI improves diagnostic precision and timing (that is, early detection). He points to clinical trials confirming, for example, that AI enhances breast cancer detection and that retinal scans can predict cardiovascular and neurodegenerative risks like Alzheimer’s and Parkinson’s disease in asymptomatic patients. Through cli-

nicians’ “liberation from the keyboard” and delegation of “data clerk” functions to AI, Topol sees a future of revitalized doctor-patient relationships, yet notes the empathy paradox that AI poses. And just as there are developments toward AI co-scientists, so too toward AI “co-clinicians” as real-time multimodal assistive tools for care teams.⁴⁷ In “The Role of AI in Drug Discovery in Africa,” Chibale argues that AI offers a unique mechanism to transition to innovation self-reliance. He advocates leveraging AI to address neglected diseases and Africa’s genetic diversity, often not included in global datasets, and using AI-driven pharmacogenomics to rectify long-standing inequities in drug efficacy. He warns, however, that without local data, infrastructure, and access, the AI era risks exacerbating the health disparities it promises to resolve.⁴⁸

In “Field Theory: AI as Social Science Question, Object & Tool,” Alondra Nelson widens the volume’s aperture. Invoking Max Weber and W. E. B. Du Bois, she frames AI not merely as an instrument but as a “social artifact” that reorganizes work, the production of knowledge, and the categories of social existence and meaning. She argues that AI poses urgent questions for the social sciences and demands a dual approach: as a methodological asset for useful, even beneficial, social inquiry and, crucially, as an object of study itself, including its effects on society. Recently, computational social scientist James Evans and colleagues framed LLMs as “cultural and social technologies” similar to the written word, economic markets, and state bureaucracies in that they do more than make information accessible; they generate “lossy but useful representations” of unmanageably large and complex human-generated data and allow it to be reorganized, transformed, and restructured in ways that create opportunities and challenges for society and for discovery.⁴⁹ Returning to Nelson, she warns that without a robust social science of AI, we risk sociotechnical and sociological failures, including “narrowing of thought” and supplanting human judgment – issues M. J. Crockett also notes in their personal brief, “Making Automation Work for Social Scientists.” Further, in their 2024 *Nature* paper with Lisa Messeri, Crockett spotlights the sociological failure of scientists overestimating their comprehension because an LLM provided them a plausible, fluent, or confident output. They also argue that the use of LLMs impacts other aspects of scientists’ work – such as reading, reviewing, and writing – risking the creation of science monocultures.⁵⁰

Arguing “Physics Is Different,” Tess Smidt elucidates the “culture and craft” of doing physics, which she describes as not a monolith but a patchwork of distinct data cultures, infrastructures, and workflows. Contrasting the faintness of gravitational waves, the industrial-scale validation in particle physics, and the artisanal data of condensed matter physics, she argues that AI-enabled science remains inextricably bound to communal norms of discovery. Smidt also reminds us of the mismatches apparent in physics, that instruments (like LIGO or synchrotrons) take decades to build, while AI advances seemingly monthly, creating temporal friction.

In her essay “Thinking & Doing Science in the Age of AI,” Alison Noble contends that AI impacts the scientific method and its workflows, transforming the scientist’s role from creator to rigorous challenger of the AI’s output.⁵¹ She distinguishes between the doing of science – increasingly AI-assisted and automated – and the thinking of science, identifying peril where cognitive offloading erodes the “habit of thinking.” Together, these perspectives offer a vital sociological counterweight to the technological views in other essays.

On the social science of discovery, two recent perspectives suggest other cautions and possibilities. Analyzing over forty million papers in the natural sciences, Qianyue Hao and colleagues reveal that while AI-enabled science expands researcher impact – tripling publications and quintupling citations compared to others – it narrows the collective aperture of inquiry, mostly because current AI capabilities direct research toward data-rich and established epistemologies over exploration of novel and data-sparse frontiers. Widening the aperture of inquiry requires “AI systems that expand not only cognitive capacity but also sensory and experimental capacity.”⁵² Elsewhere, Sendhil Mullainathan and Ashesh Rambachan argue that algorithms will reorganize science, especially the “patchwork sciences” – such as economics, medicine, and psychology – that do not converge toward a single “reigning champion” theory and rely on an arsenal of partial, often contradictory, context-dependent theories. They envision a “new factory floor” on which AI formalizes previously ineffable processes and revolutionizes discovery by learning which theories perform best and codifying the implicit judgment experts use to deploy their theoretical arsenal.⁵³

Mario Krenn and Heather Champion pose ten questions for the “Philosophy of Autonomous Science.” They posit that the (inevitable?) transition from AI tools to artificial scientists presents a profound philosophical challenge: translating epistemic traits – curiosity, surprise, creativity, taste, as well as notions of beauty or elegance – into computable, nonanthropocentric objectives. By confronting the “alien” nature of some AI discoveries and the interpretability challenges they entail, they call for a reevaluation of scientific understanding itself, questioning if human comprehension remains necessary for scientific progress. They propose moving beyond the use of AI as a tool – favored by some in this volume, along with the epistemic hierarchy that implies – toward the automation of epistemic norms. The transition is not unlike Aspuru-Guzik’s trajectory from the era of “cyborg” scientists – human-machine hybrids that enable and extend human intent – to a future of “android” AI scientists capable of generating hypotheses and, with physical embodiment, executing experiments. Such visions are dually exhilarating and disquieting and, according to Krenn and Champion, may require humans to employ “Teacher AIs” to translate concepts beyond our comprehension, lest we find ourselves like the scientists in Ted Chiang’s “Catching Crumbs from the Table,” who are reduced to hermeneutics, merely interpreting the output of metahuman AI, with humanity

losing its grasp on the fundamental workings of the world. This could bring its own anxieties in that for many scientists, the pursuit of science is fueled by epistemic emotions: curiosity, awe, surprise, and the sheer “joy of insight” and understanding. If AI recursively self-improves to the point where humans are demoted to students of its discoveries, the future of discovery could face a crisis not of capability and useful progress but of human motivation and meaning.

In “After Science” – an essay in *Science* that would sit well alongside those in this section – James Evans and philosopher Eamon Duede suggest that while there has already been a considerable amount of “AI-infused science” that has led to advances, these have been bounded by human understanding and traditional methods of justification and evaluation. But as AI becomes more capable and novelty and surprise emerge, there could be advances beyond our comprehension or ability to explain. This would be a significant shift in which AI efficacy exists independently of human legibility, thus challenging our classical notions of scientific understanding.⁵⁴ One could further question the classical scientific method and its typical starting points: For example, in “Against Theory-Motivated Experimentation,” computational cognitive scientist Marina Dubova and colleagues find that agents employing “gold-standard” or theory-driven scientific strategies – such as carefully premeditated experiments designed explicitly to falsify a dominant theory, confirm existing knowledge, or resolve a disagreement between two competing theories – create an “illusion of success,” but perform worse than agents pursuing random and undirected exploration toward novel and useful theories.⁵⁵ This suggests that the more significant value of AI systems for scientific discovery may lie not in their ability to flawlessly execute standard, theory-driven, incremental hypothesis testing, however useful, but in their capacity for high-variance, unconstrained learning from and exploration of vast data, parameter spaces, and domains that human intuition (and theories) would naturally ignore for any number of reasons (prevailing orthodoxies and reigning theories, expertise or discipline blind spots, practical infeasibilities) that could lead to novel discoveries. Given how these introductory notes opened, such shifts and futures lead some to wonder: Who gets the Nobel Prize?⁵⁶

The volume suggests other significant shifts in the future of discovery. For instance, the classic scientific method privileges empirical observation as the locus of discovery, with computation serving as a *post hoc* analytical tool. But many examples in biology and materials science suggest an ontological inversion whereby *in silico* approaches become the locus of discovery and the physical laboratory is demoted to a verification mechanism. In this way, the bottleneck for discovery shifts from the generation of ideas to the verification of ideas.⁵⁷ Here’s another: Classically, the scientific method has leaned on reductionism, the ability to derive macroscopic phenomena from fundamental microscopic laws. However, phenomena at the mesoscopic scale – such as the dynamics of a cell or the behavior of turbulent fluids – suggest the possibility that discovery may no longer require solving con-

stitutive equations (like Schrödinger's or the Navier-Stokes) but rather navigating high-dimensional latent representations that predict emergent behaviors directly from data, as LeCun and others suggest. And perhaps most daring for its implications for discovery is Hassabis's contention that if neural networks can successfully reverse-engineer and simulate stable structures of nature, from protein folding to planetary dynamics, then perhaps "information is the most fundamental unit of physics, more so than matter and energy." This would suggest that physical laws (at least the ones we have) are not immutable axioms but emergent properties of a learnable informational process. If the universe is isomorphic to a Turing machine or a neural network, then the scientific method and discovery shift from observing a static material reality to decoding a dynamic, computational one.

Perhaps one should not view such shifts as a clarion call to abandon science, as it has been, and is, productively done in ways that advance our understanding of the world. Instead, such shifts – albeit, arguably of Copernican proportions – offer ways in which the scientific enterprise could be transformed in its scope, speed, and methods, and reimaged to significantly advance discovery.

Throughout the volume, there is an underlying rumble suggesting that the future of discovery in the age of AI calls not only for deeper foundations but for a reconceptualization of how scientists are educated, including overcoming the dichotomies between theoretical, computational, and experimental science; connecting and drawing from across disciplines; having a sophisticated understanding of state-of-the-art AI capabilities (and shortcomings) and an ability to engage in new workflows and rigorously evaluate outputs; and becoming framers of questions and architects of inquiry, imbued with a sociological curiosity and imagination regarding their tools, able to critique and contribute to advances in both directions at the intersection of AI and science. The "personal briefs" by early-career scientists provide a glimpse of such future scientists. While each offers a unique perspective, they go beyond the exuberance surrounding LLMs and advocate for science-informed AI. Yet their optimism is tempered by the "frictions" of their experience and the need to preserve epistemic agency. And finally, in "When AI Meets Art at the Scales of Science," Refik Anadol uses machine learning to render the architecture of collective scientific knowledge tangibly visible, translating vast, multiscalar datasets from neuroscience, molecular biology, and astrophysics into immersive sculptures.

Where does this leave us? I hope full of enthusiasm to explore the possibilities and challenges as well as the tensions and provocations that the volume raises for the future of discovery, and to explore aspects not covered in this volume. For not only does the future of discovery matter for science, it matters for humanity's progress and prosperity.

In his 2025 Nobel Prize lecture, economic historian Joel Mokyr makes a distinction between *propositional knowledge* (science) and *prescriptive knowledge* (technology), argu-

ing that new technology drives waves of scientific advances, which in turn drive the development of new technologies, in a “positive feedback loop that can keep growing with no end in sight” – *and* it does so in ways that have historically enhanced welfare and propelled humanity’s progress.⁵⁸ The interaction of AI (prescriptive knowledge) and science (propositional knowledge) demonstrates this positive feedback loop, with two characteristics that together could further expand the possibilities for progress: AI is among a few technologies with a strong claim to being both a general-purpose technology and an invention of methods of invention.⁵⁹ In this regard, perhaps what lies ahead ought to be thought of as a combination of the Industrial Revolution (whose general purpose technologies enabled economic progress) and the Enlightenment (for the scientific method, instruments for discovery, and its “Copernican shifts”). This not only places AI in the mainstream of technological advances that have driven both profound change, and, importantly, humanity’s progress throughout history, but does so with novel and concurrent potency.

While these characteristics along with the advances highlighted and proposed in this volume project a future for discovery and societal progress full of possibilities, none of it is guaranteed or automatic – or without complexities, disruptions, and risks. On one hand, it will require developing AI that is even more capable of advancing science and enabling many more scientists in more places to participate at the expanding intersection of AI and science, not only to advance discovery in both directions but to harness it to tackle humanity’s greatest *current* and future opportunities and challenges, and benefit people everywhere – an aspect not always guaranteed, even when discovery and progress occur, as history has shown. On the other hand, it will be critically important to tackle apace the complexities of how AI reconfigures the sociological, structural and institutional, and epistemological aspects of discovery, and to take seriously the risks – including safety and ethics – that have always been at the intersection of scientific discovery and powerful technologies, taking into account the unique characteristics of AI. Both aspects will require investment in science and collaboration across academic, private-sector, and government laboratories. In addition, the issues noted at the outset that are giving rise to unease in this new era of science (and of AI generally) must be tackled so as not to impair the advancement of science, its applications, and our trust in it. All this will need active shaping by the scientific community and others outside it with focus not only on what could go wrong or be lost, but also what could go right and be gained.⁶⁰ Echoing an idea attributed to another Nobel laureate, physicist Dennis Gabor, our question *What is the future of discovery?* is not a question of prediction but of design.⁶¹

AUTHOR'S NOTE

I am grateful to the American Academy for the opportunity to conceive this *Dædalus* volume and to bring together perspectives from some of the scientists at the frontier of AI & Science. And I am grateful to the contributors. On a theme as broad as this, there are undoubtedly many more scientists and topics that are absent; for that I take responsibility. I found inspiration and learned from the contributors to this volume, from others through our conversations and collaborations—including Jeff Dean, Stuart Russell, Yoshua Bengio, Blaise Agüera y Arcas, Fei-Fei Li, Mira Murati, Dario Amodei, Dan Huttenlocher, Eric Schmidt, Erik Brynjolfsson, Nigel Shadbolt, Wendy Hall, Meghan O'Sullivan, Jennifer Doudna, Mike Spence, Michel Devoret, and Gillian Tett (especially for the access to Turing's papers and notes at King's College, Cambridge)—and from the work of many scientists, especially those cited in the endnotes. I would like to thank my colleagues at Google Research and Google DeepMind for our work together at the intersection of AI & Science, as well as colleagues at AI2o5o and Stanford's AI Index. I am especially grateful to Lizzie Dorfman (a scientist colleague whose research is also cited in this volume), who was a valuable thought-partner throughout. And as with the 2022 volume on AI & Society, this volume on AI & Science would not have come together without generous collaboration and partnership with the Academy's talented editorial team of Phyllis Bendell, Editor of *Dædalus*, who brought her experience as guide and editor and enthusiasm from the very beginning to the completion of this effort, and Peter Walton and Maya Robinson, who were creative and expert copyeditors for all the essays in this volume. These introductory notes benefited from valuable feedback from Lizzie Dorfman, Blaise Agüera y Arcas, Kerry McHugh, Kent Walker, Sarah Manyika, and Julian Manyika, but they should not be held responsible for any errors or opinions herein. The views expressed are mine and do not necessarily reflect the views of Google or Alphabet.

ABOUT THE AUTHOR

James Manyika, a Member of the American Academy since 2019, is SVP at Google-Alphabet. He is also President for Research, Labs, Tech & Society, focusing on Google and Google DeepMind's foundational and applied research and innovations in AI, computing, and science. Most recently, he served as Vice Chair of the U.S. National AI Advisory Committee established by Congress to advise the President on AI (2022–2025) and as Co-Chair of the UN Secretary General's High-level Advisory Body on AI (2023–2024). He is an inaugural Distinguished Fellow of Stanford's Human-Centered AI Institute and a Distinguished Fellow in Ethics in AI at Oxford, where he is also an appointed Visiting Professor. He is the guest editor of the Spring 2022 issue of *Dædalus* on "AI & Society."

ENDNOTES

- ¹ See Demis Hassabis, "Accelerating Scientific Discovery with AI," Nobel Prize lecture, Aula Magna, Stockholm University, December 8, 2024, <https://www.nobelprize.org/prizes/chemistry/2024/hassabis/lecture>; and David Baker, "De Novo Protein Design," Nobel

- Prize lecture, Aula Magna, Stockholm University, December 8, 2024, <https://www.nobelprize.org/prizes/chemistry/2024/baker/lecture>.
- ² For the role that AI “hallucinations” played in Baker’s work, see William J. Broad, “How Hallucinatory AI Helps Science Dream Up Big Breakthroughs,” *The New York Times*, December 23, 2024, <https://www.nytimes.com/2024/12/23/science/ai-hallucinations-science.html>. AI diffusion models such as those used by Baker are a physics-inspired deep-learning approach co-introduced in 2015 by one of the contributors to this volume, Surya Ganguli. See Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli, “Deep Unsupervised Learning Using Nonequilibrium Thermodynamics,” arXiv (2015), <https://arxiv.org/abs/1503.03585>.
- ³ For examples of AI-enabled advances beyond those in this volume, see the annual *Artificial Intelligence Index Reports* (2017–2026) by the AI Index initiative (of which I am on the steering committee) at the Stanford Institute for Human-Centered AI, available at <https://hai.stanford.edu/ai-index/2026-ai-index-report>.
- ⁴ See Rick Stevens, Valerie Taylor, Jeff Nichols, et al., *AI for Science: Report on the Department of Energy (DOE) Town Halls on Artificial Intelligence (AI) for Science* (U.S. Department of Energy Office of Scientific and Technical Information, 2019), <https://www.osti.gov/biblio/1604756>; The Royal Society, *Science in the Age of AI: How Artificial Intelligence Is Changing the Nature and Method of Scientific Research* (The Royal Society, 2024); National Academies of Sciences, Engineering, and Medicine, *Foundation Models for Scientific Discovery and Innovation: Opportunities Across the Department of Energy and the Scientific Enterprise* (National Academies Press, 2023), <https://doi.org/10.17226/29292>; National Science Foundation, “National Artificial Intelligence Research Institutes” (consisting of, at time of publication, twenty-nine AI institutes across multiple science disciplines) (accessed March 11, 2026); and U.S. Department of Energy, “Genesis Mission,” <https://genesis.energy.gov> (announced in late 2025, “A National Mission to Accelerate Science through Artificial Intelligence” involving seventeen national laboratories, academia, and the private sector).
- ⁵ See Hanchen Wang, Tianfan Fu, Yuanqi Du, et al., “Scientific Discovery in the Age of Artificial Intelligence,” *Nature* 620 (7972) (2023), <https://doi.org/10.1038/s41586-023-06221-2>.
- ⁶ See Philip Ball, “Is AI Leading to a Reproducibility Crisis in Science?” *Nature* 624 (2023): 22–24, <https://doi.org/10.1038/d41586-023-03817-6>; Arvind Narayanan and Sayash Kapoor, “Why an Overreliance on AI-Driven Modelling is Bad for Science,” *Nature* 640 (2025): 312–314, <https://doi.org/10.1038/d41586-025-01067-2>; and Editorial, “AI Scientists Are Changing Research—Institutions, Funders and Publishers Must Respond,” *Nature* 651 (2026): 853–854, <https://doi.org/10.1038/d41586-026-00934-w>. On safety, see Yoshua Bengio, Stephen Clare, Carina Prunkl, et al., *International AI Safety Report 2026* (UK Government, 2026), https://internationalaisafetyreport.org/sites/default/files/2026-02/international-ai-safety-report-2026_1.pdf; and National Academies of Sciences, Engineering, and Medicine, *The Age of AI in the Life Sciences: Benefits and Biosecurity Considerations* (National Academies Press, 2025), <https://doi.org/10.17226/28868>.
- ⁷ AI Index, *Artificial Intelligence Index Reports* (2022–2026).
- ⁸ See the reports of the UN Secretary General’s High-level Advisory Body on AI, which I cochaired: Advisory Body on Artificial Intelligence, *Interim Report: Governing AI for Humanity* (United Nations, 2023), https://www.un.org/sites/un2.un.org/files/un_ai_advisory_body_governing_ai_for_humanity_interim_report.pdf; and Advisory Body on Artificial Intelligence, *Governing AI for Humanity* (United Nations, 2024), https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf.

- ⁹ See John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon, “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence,” August 31, 1955, <http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf>. The aims of the project read today like the pursuit of artificial general intelligence: “An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.” The participants included scientists from neuroscience, cognitive science, cognitive psychology, mathematics, computer science, and information engineering. See my lecture on the seventy-fifth anniversary of the Turing Test at Kings College, Cambridge, “The Next Turing Tests,” October 15, 2025, <https://www.kingselab.org/all-events/james-manyika>. See also Eddy Keming Chen, Mikhail Belkin, Leon Bergen, and David Danks, “Does AI Already Have Human-Level Intelligence? The Evidence Is Clear,” *Nature* 650 (2026): 36–40, <https://doi.org/10.1038/d41586-026-00285-6>; and Blaise Agüera y Arcas and James Manyika, “AI Is Evolving—And Changing Our Understanding of Intelligence,” *Noema*, April 8, 2025, <https://www.noemamag.com/ai-is-evolving-and-changing-our-understanding-of-intelligence>.
- ¹⁰ Marvin Minsky, *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind* (Simon & Schuster, 2007).
- ¹¹ See President’s Information Technology Advisory Committee, *Computational Science: Ensuring America’s Competitiveness* (Executive Office of the President of the United States, 2005), https://www.nitrd.gov/pubs/pitac/pitac_report_computational-science_2005.pdf; and W. Brian Arthur, “Algorithms and the Shift in Modern Science,” Beijer Discussion Paper Series No. 269 (Beijer Institute of Ecological Economics, Royal Swedish Academy of Sciences, 2020), https://beijer.kva.se/wp-content/uploads/2020/03/Disc269_Arthur_2020.pdf. See “Theory and Observation in Science,” Stanford Encyclopedia of Philosophy.
- ¹² An early illustration of this was Stanislaw Ulam and John von Neumann’s work at Los Alamos in the 1940s that resulted in the Monte Carlo method and its use in computational physics and in Monte Carlo tree search (MCTS), formalized in 2006 and popularized by AlphaGo in 2016. See Iulia Georgescu, “The Forgotten Pioneers of Computational Physics,” *Physics World*, November 11, 2025, <https://physicsworld.com/a/the-forgotten-pioneers-of-computational-physics/>; and David Silver, Aja Huang, Chris J. Maddison, et al., “Mastering the Game of Go with Deep Neural Networks and Tree Search,” *Nature* 529 (7587) (2016): 484–489, <https://doi.org/10.1038/nature16961>. See also our recent combination of MCTS and a foundation model to solve a class of general science problems: Eser Aygün, Anastasiya Belyaeva, Gheorghe Comanici, et al., “An AI System to Help Scientists Write Expert-Level Empirical Software,” *Nature* (2026), <https://doi.org/10.1038/s41586-026-10658-6>.
- ¹³ See the chapter “The Pre-history of Computing” in Blaise Agüera y Arcas, *What Is Intelligence? Lessons from AI about Evolution, Computing, and Minds* (MIT Press, 2025).
- ¹⁴ See Alan M. Turing, “On Computable Numbers, with an Application to the Entscheidungsproblem,” *Proceedings of the London Mathematical Society* S2-42 (1) (1937): 230–265, <https://doi.org/10.1112/plms/s2-42.1.230>; Alan Turing, “‘Intelligent Machinery, A Heretical Theory,’ a Lecture Given to ‘51 Society’ at Manchester,” AMT/B/4, The Turing Digital Archive, <https://turingarchive.kings.cam.ac.uk/publications-lectures-and-talks-amtb/amt-b-4>; Alan M. Turing, “The Chemical Basis of Morphogenesis,” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 237 (641) (1952): 37–72; John von Neumann, *The Computer and the Brain* (Yale University Press, 1958); and John von Neumann, *Theory of Self-Reproducing Automata*, ed. Arthur W. Burks (University of Illinois Press, 1966).

- ¹⁵ Several prominent physicists share this view, including John Wheeler, Leonard Susskind, Paul Davies, and 2022 Nobel laureate Anton Zeilinger.
- ¹⁶ See Stephen Wolfram, “Cellular Automaton Fluids 1: Basic Theory,” *Journal of Statistical Physics* 45 (3/4) (1986). See also the chapters “Life as Computation” and “Artificial Life” in Agüera y Arcas, *What Is Intelligence?*
- ¹⁷ See Allen Newell and Herbert A. Simon, *The Logic Theory Machine: A Complex Information Processing System* (RAND Corporation, 1956). Pamela McCorduck, in “Artificial Intelligence: An Aperçu,” *Dædalus* 117 (1) (1988), tells how when the Logic Theorist proved the *Principia* theorems, Bertrand Russell (author of *The Principia Mathematica*) was delighted, but “*The Journal of Symbolic Logic* declined to publish an article—coauthored by the Logic Theorist—that described the proof,” anticipating some of today’s questions of authorship.
- ¹⁸ Margaret O. Dayhoff, Richard V. Eck, Marie A. Chang, and Minnie R. Sochard, *Atlas of Proteins Sequence and Structure* (The National Biomedical Research Foundation, 1965). See also Robert K. Lindsay, Bruce G. Buchanan, Edward A. Feigenbaum, and Joshua Lederberg, “DENDRAL: A Case Study of the First Expert System for Scientific Hypothesis Formation,” *Artificial Intelligence* 61 (1993): 209–261. For more of the history of the period from 1950–1975 and its personalities, see Pamela McCorduck, *Machines Who Think*, 2nd ed. (Routledge, 2004).
- ¹⁹ Barto and Sutton’s work, along with Geoffrey Hinton’s, influenced my early research and first publication. See Robert K. Appiah, Jean H. Daigle, James M. Manyika, and Themuso Makhurane, “Modelling and Training of Artificial Neural Networks,” *African Journal of Science and Technology Series B, Science* 6 (1) (1992); and most recently with Seijin Kobayashi, Yanick Schimpf, Maximilian Schlegel, et al., “Emergent Temporal Abstractions in Autoregressive Models Enable Hierarchical Reinforcement Learning,” arXiv (2025), <https://arxiv.org/pdf/2512.20605>. In 2025, I had the honor of presenting Andrew Barto and Rich Sutton with the Turing Award for their work on reinforcement learning at the Association of Computing Machinery (ACM) awards ceremony.
- ²⁰ Hubert L. Dreyfus and Stuart E. Dreyfus, “Making a Mind Versus Modeling the Brain: Artificial Intelligence Back at a Branchpoint,” *Dædalus* 117 (1) (Winter 1988): 15–44, https://www.amacad.org/sites/default/files/daedalus/downloads/Daedalus_Wi98_Artificial-Intelligence.pdf.
- ²¹ Jeffrey Dean, “A Golden Decade of Deep Learning: Computing Systems & Applications,” *Dædalus* 151 (2) (Spring 2022), <https://www.amacad.org/publication/daedalus/golden-decade-deep-learning-computing-systems-applications>. Foundational advances in deep learning include several by pioneers in the field whose papers are considered influential and foundational classics of this period and referred to in the field in shorthand as: AlexNet (2012), Word2Vec (2013), DQN (2013/15), VAEs (2014), GANs (2014), Adam (2014), ResNet (2015), AlphaGo (2016), Transformer (2017), BERT (2018), and GPT-3 (2020).
- ²² Kevin Scott, “I Do Not Think It Means What You Think It Means: Artificial Intelligence, Cognitive Work & Scale,” *Dædalus* 151 (2) (Spring 2022), <https://www.amacad.org/publication/daedalus/i-do-not-think-it-means-what-you-think-it-means-artificial-intelligence-cognitive-work-scale>. See Ray J. Solomonoff, “A Formal Theory of Inductive Inference, Part I,” *Information and Control* 7 (1) (1964), <https://raysolomonoff.com/publications/1964pt1.pdf>; and Ray J. Solomonoff, “A Formal Theory of Inductive Inference, Part II,” *Information and Control* 7 (2) (1964), <https://raysolomonoff.com/publications/1964pt2.pdf>. See also Universal Artificial Intelligence (AIXI) in Peter Sunehag and Marcus Hutter, “Principles of Solomonoff Induction and AIXI,” in *Algorithmic Probability and Friends. Bayesian*

- Prediction and Artificial Intelligence*, ed. David L. Dowe (Springer Nature Link, 2011). Both Solomonoff induction and AIXI have informed our work; see Alexander Meulemans, Rajai Nasser, Maciej Wolczyk, et al., “Embedded Universal Predictive Intelligence: A Coherent Framework for Multi-Agent Learning,” arXiv (2025), <https://arxiv.org/pdf/2511.22226>.
- ²³ For other examples of genomics advances in this “golden decade,” see the Telomere-to-Telomere Consortium, which decoded the final 8 percent of the human genome to produce the first complete human reference genome, using Google Research’s DeepVariant to polish the assembled sequence: Sergey Nurk, Sergey Koren, Arang Rhie, et al., “The Complete Sequence of a Human Genome,” *Science* 376 (6588) (2022), <https://doi.org/10.1126/science.abj6987>. See also the use of DeepVariant and DeepConsensus by a consortium including my colleagues to create the first draft human pangenome: Wen-Wei Liao, Mobin Asri, Jana Ebler, et al., “A Draft Human Pangenome Reference,” *Nature* 617 (7960) (2023), <https://doi.org/10.1038/s41586-023-05896-x>.
- ²⁴ See Jared Kaplan, Sam McCandlish, Tom Henighan, et al., “Scaling Laws for Neural Language Models,” arXiv (2020), <https://doi.org/10.48550/arXiv.2001.08361>; Dario Amodei, “Machines of Loving Grace,” October 2024, <https://www.darioamodei.com/essay/machines-of-loving-grace>; and Dario Amodei, “The Adolescence of Technology,” January 2026, <https://www.darioamodei.com/essay/the-adolescence-of-technology>. See also Lee Sharkey, Bilal Chughtai, Joshua Batson, et al., “Open Problems in Mechanistic Interpretability,” arXiv (2025), <https://arxiv.org/pdf/2501.16496>.
- ²⁵ “Ilya Sutskever Declares ‘Game Over for Scaling Laws’ – AI Enters New Era,” AIM Network, YouTube video, November 26, 2025, <https://www.youtube.com/watch?v=UROxY4ceHlo>. On System 1 and 2 thinking, see Daniel Kahneman, *Thinking Fast and Slow* (Farrar, Straus and Giroux, 2011).
- ²⁶ See Fei-Fei Li, “Form Words to Worlds: Spatial Intelligence is AI’s Next Frontier,” November, 2025, <https://drfeifei.substack.com/p/from-words-to-worlds-spatial-intelligence>.
- ²⁷ See Usman Anwar, Abulhair Saparov, Javier Rando, et al., “Foundational Challenges in Assuring Alignment and Safety of Large Language Models,” arXiv (2024), <https://doi.org/10.48550/arXiv.2404.09932>. See also Yoshua Bengio, Michael Cohen, Damiano Fornasiero, et al., “Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?” arXiv (2025), <https://doi.org/10.48550/arXiv.2502.15657>.
- ²⁸ Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. (Pearson, 2020). See also Stuart Russell, “If We Succeed,” *Daedalus* 151 (2) (Spring 2022): 43–57, <https://www.amacad.org/publication/daedalus/if-we-succeed>; and Stuart J. Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (Viking, 2019).
- ²⁹ Rich Sutton, “The Bitter Lesson,” Incomplete Ideas, 2019, <http://www.incompleteideas.net/InclIdeas/BitterLesson.html>.
- ³⁰ See I. J. Good’s paper “Speculations Concerning the First Ultra Intelligent Machine” (1964), in which he discusses the possibility of recursive self-improvement and the notion that such an AI would lead to an “intelligence explosion” and be “the last invention man need ever make.” He also argued that such a machine would “give the human race a good chance of surviving indefinitely” but also acknowledged “the opposite possibility, that the human race will become redundant.” See also Jack Clark, “AI Systems Are about to Start Building Themselves: The First Step Towards Recursive Self-Improvement,” Import AI 455, May 4, 2026, <https://importai.substack.com/p/import-ai-455-automating-ai-research>.

- ³¹ Another notable advance in inorganic materials is MatterGen by Microsoft Research. See Claudio Zeni, Robert Pinsler, Daniel Zügner, et al., “A Generative Model for Inorganic Materials Design,” *Nature* 639 (2025): 624–632, <https://doi.org/10.1038/s41586-025-08628-5>.
- ³² See Veronika Eyring, William D. Collins, Pierre Gentine, et al., “Pushing the Frontiers in Climate Modelling and Analysis with Machine Learning,” *Nature Climate Change* 14 (2024), <https://doi.org/10.1038/s41558-024-02095-y>. For examples of Earth foundation models, see Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, et al., “A Foundation Model for the Earth System,” *Nature* 641 (8065), <https://doi.org/10.1038/s41586-025-09005-y> (Aurora by Microsoft Research); and Daniel Klocke, Claudia Frauen, Jan Frederik, et al., “Computing the Full Earth System at 1 km Resolution,” arXiv (2025), <https://doi.org/10.48550/arXiv.2511.02021> (ICON Earth system by NVIDIA/MaxPlanck Institute). See also work with my colleagues: Aaron Bell, Amit Aides, Amr Helmy, et al., “Earth AI: Unlocking Geospatial Insights with Foundation Models and Cross-Modal Reasoning,” arXiv (2026), <https://arxiv.org/pdf/2510.18318> (Earth AI); and Christopher F. Brown, Michal R. Kazmierski, Valerie J. Pasquarella, et al., “AlphaEarth Foundations: An Embedding Field Model for Accurate and Efficient Global Mapping from Sparse Label Data,” arXiv (2025), <https://arxiv.org/abs/2507.22291v2> (AlphaEarth Foundations).
- ³³ See also Nicholas Stern (author of the influential 2006 *Stern Review on the Economics of Climate Change*) and his colleagues in Nicholas Stern, Mattia Romani, Roberta Pierfederici, et al., “Green and Intelligent: The Role of AI in the Climate Transition,” *npj Climate Action* 4 (2025), <https://doi.org/10.1038/s44168-025-00252-3>.
- ³⁴ Matthew Hutson, “AI Hunts for the Next Big Thing in Physics,” *IEEE Spectrum*, February 3, 2026, <https://spectrum.ieee.org/particle-physics-ai>. See also the recent consensus study report of the National Academies, which notes the current and future use of AI for discovery: National Academies of Sciences, Engineering, and Medicine, *Elementary Particle Physics: The Higgs and Beyond* (National Academies Press, 2025), <https://doi.org/10.17226/28839>.
- ³⁵ In “Can AI Solve Science?” Stephen Wolfram argues that while AI allows us to tackle the plethora of computationally reducible problems, the existence of problems that are computationally irreducible means that to fully advance science we must “combine the strengths of AI and of the formal computational paradigm.” See Stephen Wolfram, “Can AI Solve Science?” March 5, 2024, <https://writings.stephenwolfram.com/2024/03/can-ai-solve-science>.
- ³⁶ From Turing’s lecture to the London Mathematical Society, February 20, 1947. See also *Nature* Editorial, “Why AI Cannot Do Good Science Without Humans,” *Nature*, May 19, 2026, <https://doi.org/10.1038/d41586-026-01551-3>. For an example, see Syed Asad Rizvi, Daniel Levine, Aakash Patel, et al., “Scaling Large Language Models for Next Generation Single-Cell Analysis,” bioRxiv (2025), <https://doi.org/10.1101/2025.04.14.648850>, by Yale and Google Research.
- ³⁷ For example, see the use of AlphaTensor to break a fifty-year record in matrix multiplication using tensor decomposition as a reinforcement learning game in Alhussein Fawzi, Matej Balog, Aja Huang, et al., “Discovering Faster Matrix Multiplication Algorithms with Reinforcement Learning,” *Nature* 610 (7930) (2022), <https://doi.org/10.1038/s41586-022-05172-4>. In hashing and sorting, see AlphaDev in Daniel J. Mankowitz, Andrea Michi, Anton Zhernov, et al., “Faster Sorting Algorithms Discovered Using Deep Reinforcement Learning,” *Nature* 618 (7964) (2023), <https://doi.org/10.1038/s41586-023-06004-9>. In NP-Hard combinatorics, see FunSearch, which uses LLMs as

- intelligent mutation engines embedded within an evolutionary algorithm, in Bernardino Romera-Paredes, Mohammadamin Barekatin, Alexander Novikov, et al., “Mathematical Discoveries from Program Search with Large Language Models,” *Nature* 625 (2024), <https://doi.org/10.1038/s41586-023-06924-6>.
- ³⁸ See Bogdan Georgiev, Javier Gómez-Serrano, Terence Tao, and Adam Zsolt Wagner, “Mathematical Exploration and Discovery at Scale,” arXiv (2025), <https://doi.org/10.48550/arXiv.2511.02864>; and Romera-Paredes, Barekatin, Novikov, et al., “Mathematical Discoveries from Program Search with Large Language Models.” See also Benjamin Skuse, “AI Is Acing Math Exams Faster than Scientists Write Them,” *IEEE Spectrum*, February 25, 2026, <https://spectrum.ieee.org/ai-math-benchmarks>.
- ³⁹ For other recent genetics advances, see Nauman Javed, Thomas Weingarten, Arijit Sehano-bish, et al., “A Multi-Modal Transformer for Cell Type-Agnostic Regulatory Predictions,” *Cell Genomics* 5 (2) (2025); Žiga Avsec, Natasha Latysheva, Jun Cheng, et al., “Advancing Regulatory Variant Effect Prediction with AlphaGenome,” *Nature* 649 (8099) (2026), <https://doi.org/10.1038/s41586-025-10014-0>; and Garyk Brix, Matthew G. Durrant, Jerome Ku, et al., “Genome Modelling and Design across All Domains of Life with Evo 2,” *Nature* 652 (8112) (2026), <https://doi.org/10.1038/s41586-026-10176-5>.
- ⁴⁰ See papers coauthored with my colleagues and our collaborators on quantum and related science advances: Google Quantum AI and Collaborators, “Quantum Error Correction Below the Surface Code Threshold,” *Nature* 638 (8052) (2025), <https://doi.org/10.1038/s41586-024-08449-y>; Google Quantum AI and Collaborators, “Observation of Constructive Interference at the Edge of Quantum Ergodicity,” *Nature* 646 (8086) (2025), <https://doi.org/10.1038/s41586-025-09526-6>; C. Zhang, R. G. Cortiñas, A. H. Karamlou, et al., “Quantum Computation of Molecular Geometry via Many-Body Nuclear Spin Echoes,” arXiv (2025), <https://doi.org/10.48550/arXiv.2510.19550>; and Charina Chou, James Manyika, and Hartmut Neven, “The Race to Lead the Quantum Future,” *Foreign Affairs*, January/February 2025, <https://www.foreignaffairs.com/united-states/race-lead-quantum-future-chou-manyika-neven>.
- ⁴¹ See also Mohammad Ghazi Vakili, Christoph Gorgulla, Jamie Snider, et al., “Quantum-Computing-Enhanced Algorithm Unveils Potential KRAS Inhibitors,” *Nature Biotechnology* 43 (12) (2025), <https://doi.org/10.1038/s41587-024-02526-3>.
- ⁴² For examples of such tools, see AI Scientist in Chris Lu, Cong Lu, Robert Tjarko Lange, et al., “Towards End-to-End Automation of AI Research” *Nature* 651 (2026): 914–919, <https://doi.org/10.1038/s41586-026-10265-5>; CORAL in Ao Qu, Han Zheng, Zijian Zhou, et al., “CORAL: Towards Autonomous Multi-Agent Evolution for Open-Ended Discovery,” arXiv (2026), <https://doi.org/10.48550/arXiv.2604.01658>; and SAGA in Yuanqi Du, Botao Yu, Tianyu Liu, et al., “Accelerating Scientific Discovery with Autonomous Goal-Evolving Agents,” arXiv (2026), <https://doi.org/10.48550/arXiv.2512.21782>. See also our work in Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, et al., “Accelerating Scientific Discovery with Co-Scientist,” *Nature* (2026), <https://doi.org/10.1038/s41586-026-10644-y>.
- ⁴³ Judea Pearl and Dana Mackenzie, *The Book of Why: The New Science of Cause and Effect* (Basic Books, 2020).
- ⁴⁴ See Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, et al., “Towards Causal Representation Learning,” <https://arxiv.org/abs/2102.11107>; and Goutham Rajendran, Simon Buchholz, Bryon Aragam, et al., “From Causal to Concept-Based Representation

- Learning,” in *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, ed. Amir Globerson, Lester Mackey, Danielle Belgrave, et al. (Curran Associates, Inc., 2024).
- ⁴⁵ See Silviu-Marian Udrescu and Max Tegmark, “AI Feynman : A Physics-Inspired Method for Symbolic Regression,” *Science Advances* 6 (16) (2020).
- ⁴⁶ See Sebastian Musslick, Laura K. Bartlett, Suyog H. Chandramouli, et al., “Automating the Practice of Science : Opportunities, Challenges, and Implications,” *Proceedings of the National Academy of Sciences* 122 (5) (2025), <https://doi.org/10.1073/pnas.2401238121>. See also, from 1996, Gillies, *AI and the Scientific Method*; and his 2022 update given AI progress, Donald Gillies, “Artificial Intelligence and Philosophy of Science from 1990s to 2020,” in *Current Trends in Philosophy of Science: A Prospective for the Near Future*, ed. Wenceslao J. Gonzalez (Synthese Library, 2022).
- ⁴⁷ See recent results from field studies in breast cancer detection in collaborations between Imperial College London, the UK’s National Health Service (NHS), and Google Research : Lucy M. Warren, Jenny Venton, Kenneth C. Young, et al., “Impact of Using Artificial Intelligence as a Second Reader in Breast Screening Including Arbitration,” *Nature Cancer* 7 (2026), <https://www.nature.com/articles/s43018-026-01128-z>; and Christopher J. Kelly, Marc Wilson, Lucy M. Warren, et al., “Diagnostic Accuracy, Fairness and Clinical Implementation of AI for Breast Cancer Screening : Results of Multicenter Retrospective and Prospective Technical Feasibility Studies,” *Nature Cancer* 7 (2026), <https://www.nature.com/articles/s43018-026-01127-0>. On tools for clinicians, see Derek C. Angus, Rohan Khera, Tracy Lieu, et al., “AI, Health, and Health Care Today and Tomorrow : The JAMA Summit Report on Artificial Intelligence,” *JAMA* 334 (18) (2025). See work with my colleagues : Khaled Saab, Chunjong Park, Tim Strother, et al., “Advancing Conversational Diagnostic AI with Multimodal Reasoning,” *Nature Medicine* (2026), <https://doi.org/10.1038/s41591-026-04371-0>; and Elahe Vedadi, David Barrett, Natalie Harris, et al., “Towards Physician-Centered Oversight of Conversational Diagnostic AI,” arXiv (2025), <https://doi.org/10.48550/arXiv.2507.15743> (forthcoming in *Nature*). See also David Autor and James Manyika, “A Better Way to Think About AI,” *The Atlantic*, August 24, 2025, <https://www.theatlantic.com/technology/archive/2025/08/ai-job-loss-human-enhancement-google/683963>.
- ⁴⁸ Georgia Channing and Avijit Ghosh, “AI for Scientific Discovery Is a Social Problem,” *Patterns* 7 (3) (2026), <https://doi.org/10.1016/j.patter.2026.101497>.
- ⁴⁹ Henry Farrell, Alison Gopnik, Cosma Shalizi, and James Evans, “Large AI Models are Cultural and Social Technologies,” *Science* 387 (6739) (2025), <https://doi.org/10.1126/science.adt9819>.
- ⁵⁰ See Lisa Messeri and M. J. Crockett, “Artificial Intelligence and Illusions of Understanding in Scientific Research,” *Nature* 627 (8002) (2024), <https://doi.org/10.1038/s41586-024-07146-0>; and Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, et al., “A Large-Scale Randomized Study of Large Language Model Feedback in Peer Review,” *Nature Machine Intelligence* (2026), <https://doi.org/10.1038/s42256-026-01188-x>.
- ⁵¹ In David P. Woodruff, Vincent Cohen-Addad, Lalit Jain, et al., “Accelerating Scientific Research with Gemini: Case Studies and Common Techniques,” arXiv (2026), <https://doi.org/10.48550/arXiv.2602.03837>, we tackle, with academic collaborators, several open problems in computer science, physics, optimization, and economics with some novel results and highlight new workflows and the necessity of human orchestration. See also Don Knuth, “Claude’s Cycles” (2026), <https://www-cs-faculty.stanford.edu/~knuth/papers/claude-cycles.pdf>, in which Knuth (Turing Award winner, considered the founding father of the rigorous analysis of algorithms) and his colleagues

use Claude to solve an open combinatorial problem about directed Hamiltonian cycles that he had been stuck on.

- ⁵² Qianyue Hao, Fengli Xu, Yong Li, and James Evans, “Artificial Intelligence Tools Expand Scientists’ Impact but Contract Science’s Focus,” *Nature* 649 (8099) (2026), <https://doi.org/10.1038/s41586-025-09922-y>. See also Sabina Leonelli and Alexander Mussnug, “Convenience AI,” *PhilSci Archive* (2025), <https://philsci-archive.pitt.edu/24891>.
- ⁵³ Sendhil Mullainathan and Ashesh Rambachan, “Science in the Age of Algorithms,” in *The Economics of Transformative AI*, ed. Ajay K. Agrawal, Erik Brynjolfsson, and Anton Korinek (forthcoming with University of Chicago Press). See also Nicola Jones, “Half of Social-Science Studies Fail Replication Test in Years-Long Project,” *Nature*, April 1, 2026, <https://doi.org/10.1038/d41586-026-00955-5>. For an example of AI use in the historical sciences, see Donal Khosrowi and Finola Finn, “Can Generative AI Produce Novel Evidence?” *Philosophy of Science* 92 (5) (2025), <https://doi.org/10.1017/psa.2025.10123>.
- ⁵⁴ James Evans and Eamon Duede, “After Science,” *Science* 390 (6774) (2025), <https://doi.org/10.1126/science.aec7650>.
- ⁵⁵ Marina Dubova, Arseny Moskvichev, and Kevin Zollman, “Against Theory-Motivated Experimentation in Science,” *MetaArXiv* (2025), https://doi.org/10.31222/osf.io/ysv2u_v2.
- ⁵⁶ See Dashun Wang, “Prizes Must Recognize Machine Contributions to Discovery,” *Nature* 646 (2025), <https://doi.org/10.1038/d41586-025-03217-y>.
- ⁵⁷ A defining characteristic of late-twentieth-century philosophy of science is the dissolution of the strict dichotomy between theory and observation, with many embracing the “theory-ladenness” of observation—the realization that empirical results are inherently the result of prior theoretical commitments and values, and the limitations of the instruments and signal-processing algorithms used to produce them. See “Theory and Observation in Science,” *Stanford Encyclopedia of Philosophy*, rev. January 12, 2026, <https://plato.stanford.edu/entries/science-theory-observation>. See also discussions on the philosophical status of simulations and *in silico* experiments in “Computer Simulations in Science,” *Stanford Encyclopedia of Philosophy*, rev. February 19, 2026, <https://plato.stanford.edu/entries/simulations-science>.
- ⁵⁸ See Joel Mokyr, “The Past and Future of Innovation : Can Progress Be Sustained?” Nobel Prize lecture, Aula Magna, Stockholm University, December 8, 2025, <https://www.nobelprize.org/prizes/economic-sciences/2025/mokyr/lecture>.
- ⁵⁹ See Martin Neil Bailly, David M. Byrne, Aidan T. Kane, and Paul E. Soto, “Generative AI at the Crossroads: Light Bulb, Dynamo, or Microscope?” (Brookings Institution, 2025), https://www.brookings.edu/wp-content/uploads/08/2025/bailly-byrne-kane-soto_generative-ai-crossroads_2025-09-05_FINAL-WEB.pdf; and Nicholas Crafts, “Artificial Intelligence as a General-Purpose Technology : An Historical Perspective,” *Oxford Review of Economic Policy* 37 (3) (2021), <https://doi.org/10.1093/oxrep/grab012>.
- ⁶⁰ On the need for academic and private sector collaboration for AI for science, see James Manyika, “Bidirectional Collaboration,” part of Marcia McNutt, “For a More Competitive U.S. Research Enterprise, the Work Begins Now,” *Issues in Science and Technology* 41 (1) (2024), <https://issues.org/state-of-the-science-mcnutt>. See also James Manyika, “Getting AI Right: A 2050 Thought Experiment,” in *The Digitalist Papers: Artificial Intelligence and Democracy in America* (Stanford Digital Economy Lab, 2024), <https://www.digitalistpapers.com/essays/getting-ai-right>.
- ⁶¹ Dennis Gabor, *Inventing the Future* (Secker & Warburg, 1963).

AI as the Ultimate Tool for Science: A Conversation with Demis Hassabis

James Manyika. Demis Hassabis has focused on artificial intelligence from an early age. With a background in computer science and cognitive neuroscience, Demis first garnered wide attention for his work on AlphaGo, an AI system that in 2016 defeated a world champion in the complex strategy game of Go, long considered a grand challenge for machines playing games. But his ultimate goal was always much bigger: to use AI to solve problems in science and understand the deepest mysteries of the universe. His first proof point for this vision was AlphaFold, an AI system that cracked the fifty-year-old protein structure prediction problem. He was awarded the 2024 Nobel Prize in Chemistry and the 2023 Breakthrough Prize in Life Sciences for this work. Demis has also long been a pioneering entrepreneur in AI, most notably as cofounder and CEO of DeepMind, one of the AI labs leading the development of frontier foundation models. Fundamentally, Demis thinks of himself as a scientist, and so this dialogue focuses on AI and science: what motivates him as a scientist working with AI; what progress, opportunities, and challenges he sees; and more.

Manyika. Why does science matter to you?

Demis Hassabis. I've always been driven by a deep desire to better understand the world around us. The scientific method is arguably the most important idea humanity has ever had to make progress and advance knowledge. For me, there's nothing more thrilling than being at the absolute frontier of what's known, pushing the boundaries forward and discovering something new for the very first time – that's science at its very best.

Manyika. Why do you think of yourself as a scientist, and how did you get started as a scientist?

Hassabis. Today, I think of myself first and foremost as a scientist. But when I started off my career programming AI for games and studying computer science as an undergraduate at Cambridge, I thought of myself as an engineer. Then as I followed up on my long-standing interest in the brain and started studying cognitive neuroscience, and specifically memory and imagination for my PhD, I realized I worked best in the mode of a scientist coming up with hypotheses about the world and experiments to prove or disprove them.

But I'm also very practical and interested in real-world impact. That's where the engineering half of my brain still comes in. Instead of getting lost in a theoretical argument, I lean into my pragmatic side and start thinking about how to build tools or experiments to prove hypotheses. AlphaGo and AlphaFold are good examples of that.¹

Manyika. What scientific accomplishments have inspired you the most?

Hassabis. I'm a huge admirer and fan of the polymaths who lived during the Renaissance and Classical Greek eras, the greats like Aristotle and da Vinci, because they didn't see any boundaries between subjects. I love that because it allows one to find connections between somewhat disparate domains. I believe those combinations are where the biggest advances will come from in the next ten to twenty years, and I think AI is going to help by finding commonalities or crossover between fields.

I'm also inspired by enormous leaps of imagination and intuition that, looking back through the prism of time, come seemingly out of the blue. Einstein, Newton, Bohr, Ramanujan, and Feynman are some of my favorites – the scientists and mathematicians who have a mix of ideas, expertise, and interests going around in their minds and then make an intuitive leap in terms of how the world works.

Manyika. What is on your list of the most important scientific problems that we haven't solved yet?

Hassabis. I got into science and AI because as a kid I was obsessed with, and I suppose in some way you could say haunted by, the biggest questions about the nature of reality and consciousness that underpin our understanding of the world. What is time? What is gravity? How does it all fit together with the quantum world? Amazingly, there's not been a lot of progress on some of those fundamental questions in the last fifty-plus years, in part because they're so complex.

I've always seen AI as potentially the ultimate tool to advance the frontier of knowledge. Whether by processing data or assisting us in coming up with new conjectures, I think AI could help us tackle these fundamental questions.

Manyika. How do you link these questions about the nature of reality with a computational approach to understanding them?

Hassabis. I've always been fascinated by this question of what can and can't be modeled by a classical computer or a Turing machine, and what the answer would tell us about the nature of reality. Most things in nature have a stable structure that evolved and survived over time – the behavior of a cell, the shape of a mountain, even the orbit of a planet. That suggests there should be some sort of pattern or process that is potentially learnable by a neural network. AI systems like Alpha-

Fold are already showing us we can model many more complex systems with classical computers than we previously thought.

Maybe in the limit most natural systems can be simulated or learned – kind of reverse engineered – by a classical system. That would support a conjecture I have that information is the most fundamental unit of physics, more so than matter and energy, and is the best way to understand the true nature of the universe.

Manyika. In what way do you think information may be the most foundational aspect of understanding reality? If it is, how could a computational approach to understanding information help us answer the foundational questions?

Hassabis. Traditionally in physics, energy and matter are regarded as primary. We can observe and measure them. Then people like Claude Shannon, Alan Turing, and John von Neumann – the founding fathers of computer science – introduced information entropy, measures of information, mutual information, and other important related concepts. But they came after the physicists, so they were trying to fit information theory within the existing framework of physics. Maybe it should have been the other way around, and physics should have been framed in terms of information.

This notion also relates to Turing machines, of course, and what kinds of processes are computable on classical systems. Physicists assume that the universe is a quantum system, and that classical systems can't model quantum systems due to their enormous complexity. But if it turns out a lot more physical processes in nature are actually able to be accurately modeled by classical systems, it may have some important implications for our understanding of physics.

Ultimately, I'm most interested in the nature of reality, and I question whether our current understanding of quantum systems is the whole story. I suspect we might eventually find a simpler underlying description of what we are seeing that turns out to be more amenable to being modeled by a classical system, and I think we're more likely to discover that if we think about the world as a computational universe.

Manyika. Let's take that a step further. If you take an informational or computability view of the world, don't we run into difficulties, such as Kolmogorov complexity, the P equals NP question of whether problems are solvable in a tractable amount of time, or the Halting problem in Turing machines? Does that not potentially place a limit on how far we can go with the computational approach?

Hassabis. It does, and I think those are the interesting limits to test and understand. P equals NP – which attempts to categorize the difficulty of a problem by how much computation it would take to find and check a solution, respectively – is one of the most important questions in science to resolve. I suspect P is not equal to NP , and there are some problems out there that are just not tractable to solve in

a practical amount of time without invoking the help of, say, a quantum computer, but we need to understand this a lot better because there may be more nuance here than we previously realized. In our work with AlphaGo and AlphaFold, we're showing that if you do a lot of precompute, which is not normally considered in these kinds of scenarios, you can seemingly answer some highly complex questions approximately optimally in P (polynomial) time. Neural networks are effectively using massive amounts of precompute to compress knowledge into some efficient artifact. That computed artifact is then available at test time and, for a lot of natural systems, you can use it to narrow down your search space so you don't have to consider all the possible configurations they could potentially take, but only a much smaller subset that are actually plausible.

Let's take proteins. There are roughly 10^{300} possible conformations of an average protein. It would take longer than the age of the universe to enumerate that exhaustively to find the one specific shape it takes, so you have to do something much smarter. You have to learn what patterns there are for different amino acid sequences and then only search a tiny fraction of the possibilities to find the approximately correct solution. That seems to be what we managed to do with AlphaFold. Maybe not perfectly, but to an approximation that is at least good enough for practical purposes.

Similarly, there are 10^{170} possible positions in the game of Go; there is no way to find the best move in a position through sheer brute force computation using simple heuristics. Instead, AlphaGo did this by learning a good model of plausible Go moves, so it only needed to consider the most fruitful paths of moves, and then from that much smaller number of options find the best move that would most likely get it to a winning position.

There's something very interesting going on here in successfully navigating these enormous search spaces, and my hunch is it's pretty profound. We have a front row seat to these fundamental questions because of what we're seeing in AI. Not just with AlphaGo, AlphaFold, and our science projects, but also with our video models like Veo 3. They are modeling some aspects of reality just by watching a few billion videos. It seems pretty amazing that they can reverse engineer some kind of intuitive physics: how liquids flow, reflections in glass, or the shadows cast by objects.

All of this is telling us something about the underlying nature of physics, including what the limit of a Turing machine is versus what you need a quantum system for, and, therefore, something about the nature of reality itself. AI has implications for all those questions.

Manyika. AlphaFold seems to be a great example of what you've said about the natural world having structure that can be modeled or learned by neural networks. Talk about how you did that in the case of AlphaFold.

Hassabis. AlphaFold was our solution to the protein folding or protein structure prediction problem. You start with an amino acid sequence – you can think of it very roughly as the genetic sequence for the protein, a one-dimensional string of letters. In the body or in nature, that string folds up into a 3D structure, and that shape goes a long way toward defining the function of that protein, which is really important for drug discovery and disease understanding.

Predicting those structures is a fifty-year-old grand challenge in biology. We managed to solve it with AlphaFold, at least to an atomic accuracy, which is the level of precision needed for experimentalists to find it useful.

The way we did it is that there were about 150,000 known structures that had been painstakingly put together by structural biologists over the past thirty to forty years with very expensive equipment like electron microscopes. That was just about enough data to give our AI system clues as to the topology of proteins. Of course they don't just fold up randomly; there are some constraints, and the AI system learned them. Eventually it was able, within a few seconds, to come up with a plausible structure for an unseen protein. Then over the course of a year we folded all two hundred million proteins known to science and, with our colleagues at the European Bioinformatics Institute, made them freely accessible in a vast database for researchers all around the world to use.

Manyika. Of course, you've extended this to AlphaFold 3 and actual protein interactions. Are cells next?

Hassabis. AlphaFold 2 essentially predicted the static snapshot structure of a protein. But we know biology is a dynamic system and we need to understand protein interactions and how they change their shape and behavior.

AlphaFold 3 is our next step on the ladder.² We're now predicting all pairwise interactions between proteins and proteins, proteins and ligands, proteins and RNA, proteins and DNA. Basically, all of life's molecules and their interactions in pairs can be modeled by AlphaFold 3.

But to scale up to an entire cell, we're going to need to do much better than that. There are cascades of interactions and more than two biomolecules are involved. It is obviously hugely complex. The next stage might be modeling a pathway and then maybe a simple cell, like a yeast cell, which is an entire organism. Then, if that works, maybe we simulate an organelle from a more complex organism, and then eventually an entire organism.

I've had this dream for over twenty years. The idea would be to have a realistic enough simulation of a cell that you could do *in silico* experiments maybe orders of magnitude faster than you could do in the wet lab. Then the lab is saved for the validation step and the grounding of the results rather than for the exploration phase. A virtual cell would be an incredible boon not only for fundamental biology but also for drug discovery.

Manyika. You have also taken on other areas in science – physics, chemistry, weather, climate systems, and so forth. Is the common feature there that these are all natural systems, and therefore they lend themselves to being learned?

Hassabis. We look for three aspects of a problem in determining whether it is suitable to tackle with the AI techniques we have today. First, can the problem be described as or converted into a description of a massive combinatorial space? Perhaps it's intractably large and normal brute force techniques won't work. Second, if that's true, do you have enough data to learn some sort of model of the topology of that space? Or maybe a simulator is available or learnable that can generate some additional synthetic data. Ideally, you have both. Third, you need a clear objective that you're trying to minimize or maximize. In games, that is winning or maximizing the score. In a natural system, that might be minimizing the free energy in that system. If you can quantify that, you can then use a model to search with the guidance of the objective function toward the optimal solution. It turns out a lot of problems fit that description.

Beyond that, we look for problems where the impact of solving it would be huge. If you think about the tree of all knowledge, there are some root node problems that, if you solve them, you unlock entire new branches of research. Protein folding was definitely one and it opens up opportunities in drug discovery and disease understanding. Whether fusion for energy, weather systems, material design, or fundamental mathematics – all of the problems we try to work on have those characteristics.³

Manyika. Say more about math. It is not intuitively obvious that all of mathematics has structure, at least not structure we understand. Distribution of prime numbers is perhaps a good example.

Hassabis. There's obviously still an open question about how much structure there is in numbers. I suspect it is quite a lot because math was developed to describe the real world, which does have structure, and it seems to be able to do that very well.

There are brilliant mathematicians like Srinivasa Ramanujan, who was incredibly intuitive and came up with amazing solutions without going through all the intermediate steps, so much so that he felt these leaps of imagination were divinely inspired. We call it implicit knowledge or intuition because we can't consciously access it or explicitly explain it in a structured way.

But intuition is not magic; it is just the word we use for the implicit knowledge the brain acquires through experiences. People who are very good at certain domains, say, like music, have extremely efficient and highly abstract versions of that knowledge. There's still a process there though, and I think in the limit it could be modeled by an AI system.

We're applying AI to mathematics and already seeing a glimpse of what can be modeled. Our recent Gemini model, Deep Think, reached gold medal-level performance in the International Mathematical Olympiad, which is a super difficult math challenge for the top high school students in the world.⁴ The model is given a natural-language problem that it converts into a formal mathematical language, then solves it in an iterative way by performing some mathematical operations on the current working version of a proof until it reaches a verifiably correct proof.

But that's quite different from understanding if there are patterns in the numbers or structure in mathematics. That would require another kind of system. I think both are interesting – AI solving mathematical conjectures and finding deeper patterns in mathematics – and I think both may be possible.

Manyika. What gaps and limitations do you see in current AI systems when it comes to being able to do science? What new breakthroughs do you think are needed?

Hassabis. One thing that's missing and one of the reasons why I don't think today's systems are close to AGI (artificial general intelligence) yet is that they don't have true creativity. What I mean by that is, can you have an AI system that doesn't just solve a difficult problem, which is obviously quite useful and impressive, but can actually come up with a beautiful or meaningful conjecture like the Millennium Prize problems or the Riemann hypothesis? I think that's much more difficult.

Asking the right question is the hardest part of science. What's the right hypothesis that's worth exploring and worth doing science on? We don't fully understand how that level of creativity works in human experts, and we certainly don't have it yet in our AI systems.

The other missing feature in current AI systems is consistency. On the one hand, these systems can solve very complicated Math Olympiad problems; on the other hand, if you pose the question slightly differently, it can still mess up on elementary math. That wouldn't be the case for a human mathematician. They would have consistency of capability across the field.

Manyika. What does the idea of AGI mean to you?

Hassabis. I've always defined AGI as a system that exhibits all the cognitive capabilities the human mind has. That's a critical reference point because the human brain is the only existence proof that we know of so far, perhaps in the universe, that general intelligence is even possible. The brain is extremely general and adaptable – it's incredible that with minds evolved for hunter-gathering, we were able to invent the scientific method and build the marvels of modern civilization around us, from telescopes to jets to computers. The ingenuity of history's great artists, scientists, and philosophers all emerged from this same remarkable brain architecture. In my opinion, we won't really know for sure that we've built

a fully general system until we're able to match the entirety of what we know that brain architecture can do.

From a theoretical point of view, this definition is closely related to the concept of a Turing machine, which Alan Turing famously conjectured could compute any algorithm given infinite time and memory. The brain can be viewed as a biological approximation to a Turing machine, meaning that, in theory, it could learn anything that is computable. However, as with any finite system, there always has to be some degree of specialization because there is only so much time, memory, and information. No practical system can be optimized for every possible task simultaneously; we can't circumvent the "no free lunch" theorem. The key, though, is the brain's potential to learn almost anything, especially as we are able to design and build tools and machines to help us gain knowledge.

Any candidate AGI system should have the consistency across the board to cover the incredible range of human abilities, including being capable of true intuitive leaps of discovery and coming up with a useful novel hypothesis or conjecture. This could be tested by, for example, training an AI system with a knowledge cutoff date of 1910 and seeing if it is able to come up with General Relativity like Einstein famously did in 1915. For now, the answer is clearly no.

What foundation models can already do today is extremely impressive, and in my opinion there's no doubt that they will ultimately form the backbone of any future AGI system. Whether or not scaling up these models further is enough on its own to reach AGI is an open empirical question, and I would not be surprised if a small number of breakthroughs on the level of transformers or deep reinforcement learning were still needed. So that's why at Google DeepMind we continue to relentlessly push forward both the frontiers of scaling and innovation in parallel.

Manyika. In assessing whether we've achieved AGI, on the one hand, one could use a battery of tests, and, on the other hand, perhaps there is some intuitive leap or spark that will happen – maybe like move thirty-seven (when AlphaGo went outside human strategy and intuition to play an incredibly unlikely move that led to victory over world champion Go player Lee Sedol) – and we'll say, "Ah, I think we've got it!" Is it both, or is it one or the other?

Hassabis. I think you need both. This is such an important moment in human history and I think we should be really scientific and sure about it. There should be quite a lot of obvious examples of invention or creativity happening, where some new physics is invented or some new conjecture is proposed that mathematicians agree is super meaningful. Instead of just coming up with a great new strategy in Go like AlphaGo did, could it invent a game as elegant as Go? I would want to see several examples of this happening for it to at least start even being a candidate.

Then I think we should have this exhaustive scientific approach to make sure there aren't any flaws or holes in the system. An example of that today is LLMs

(large language models). It's very easy for the average person to come up with things they can't do. For example, they can't even play chess to an amateur level or learn a new game, and they don't continually learn. There are problems with reasoning, memory, and consistency. All of those would need to be solved in order to pass a test like the one I was proposing.

Manyika. How should scientists think about AI when it comes to discovery? Is AI a tool, a collaborator, something altogether different? And how do you imagine the role of scientists in the future when we have these more capable systems?

Hassabis. For the next decade or so, we should think about AI as this amazing tool to help scientists, whether that's number crunching, pattern matching, or some sort of low-level coding or engineering. Used in these ways, I think it is going to be the ultimate tool to help advance science and medicine and it will usher in a new golden era of discovery. Conjectures and hypotheses are going to come from the human experts, and then they'll have these tools that make them much more efficient – maybe ten times or one hundred times more – at what they do. Beyond that timeframe, it is hard to say with any certainty, but perhaps these systems will become more like collaborators.

Manyika. What does it mean to you to use AI in science responsibly? I know you care deeply about this, but what does that actually mean?

Hassabis. It means adhering to the scientific method and being extremely rigorous and careful with your tests and the claims made about the systems. It also means being very thoughtful about what you deploy and use the systems for and trying to think through their second-order effects. You're never going to get that perfect because it's such a new technology and the frontier is evolving so fast, but one should make every possible effort.

Related to that, it means using the scientific method itself to help with understanding AI: its limitations, its capabilities, and its risks. That understanding should then be used to put the right guardrails on AI, which are then monitored and studied as well.

I think the scientific method is the best approach to tackling the challenges with AI as it becomes more prevalent and more powerful, rather than perhaps a more hacker mentality of building first and asking questions later. That has its place, of course, and it's amazing for progress, but with something as transformative and as consequential as AI, we need to use the best tools we have, which in my opinion is the scientific method, to have as much foresight and understanding of these technologies as we can.

Manyika. What are the biggest benefits to humanity that you'd like to see from your work?

Hassabis. Number one is improving human health and curing diseases, especially those that disproportionately affect the poorer parts of the world. I think we can do so much better than we're doing today.

Then there is helping with energy, climate, and the environment by developing new materials and technologies faster, optimizing existing infrastructure and grids, and also modeling what's happening to the world and how we're affecting ecosystems. In my opinion, those are the most important things we can use AI for in the near term.

Manyika. If we're having this conversation in 2050, what are we going to be talking about?

Hassabis. Given the pace of how fast things are advancing, almost week by week, the only thing we know for certain is that by 2050 things will likely look very different. Hopefully the field will have safely stewarded AGI into the world for the benefit of all humanity, and we will be discussing how to navigate a post-AGI world of radical abundance and living in a new golden era of scientific discovery and wonder!

ABOUT THE AUTHORS

Demis Hassabis, a Member of the American Academy since 2021, is a Nobel laureate and an artificial intelligence researcher. He is Co-Founder and CEO of DeepMind and of Isomorphic Labs. In 2016, DeepMind's AlphaGo became the first program to beat a world champion at the complex game of Go. In 2020, its AlphaFold 2 program solved the fifty-year grand challenge of protein structure prediction.

James Manyika, a Member of the American Academy since 2019, is SVP at Google-Alphabet and President for Research, Labs, Tech & Society.

ENDNOTES

- ¹ David Silver, Aja Huang, Chris J. Maddison, et al., “Mastering the Game of Go with Deep Neural Networks and Tree Search,” *Nature* 529 (7587) (2016): 484–489; and John Jumper, Richard Evans, Alexander Pritzel, et al., “Highly Accurate Protein Structure Prediction with AlphaFold,” *Nature* 596 (7873) (2021): 583–589.
- ² Josh Abramson, Jonas Adler, Jack Dunger, et al., “Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3,” *Nature* 630 (8016) (2024): 493–500.
- ³ Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, et al., “Probabilistic Weather Forecasting with Machine Learning,” *Nature* 637 (8044) (2025): 84–90; and Yongji Wang and Sam Blackwell, “Discovering New Solutions to Century-Old Problems in Fluid Dynamics,” Google DeepMind, September 18, 2025, <https://deepmind.google/blog/discovering-new-solutions-to-century-old-problems-in-fluid-dynamics>.
- ⁴ Thomas Hubert, Rishi Mehta, Laurent Sartran, et al., “Olympiad-Level Formal Mathematical Reasoning with Reinforcement Learning,” *Nature*, November 12, 2025 (unedited, prepublication manuscript), <https://www.nature.com/articles/s41586-025-09833-y>.

Learning Abstractions: A Conversation with Yann LeCun

James Manyika. Yann LeCun is widely considered one of the godfathers of the modern era of artificial intelligence. His pioneering research with Geoffrey Hinton and Yoshua Bengio on deep learning won them the 2018 Turing Award, considered the Nobel Prize for computer science, and in 2025, the three were awarded the Queen Elizabeth Prize for Engineering. Yann has long focused on foundational and scientific advances in AI, from biologically inspired convolutional neural networks and graph transformer networks to energy-based models for machine learning and world models. In this dialogue, we focus on foundational advances in AI, what he sees as progress to date, the limitations of current mainstream approaches, and what's needed to further advance AI and benefit science, including the role of scientists and the importance of open science.

Manyika. The last decade in artificial intelligence has been extraordinary. How would you characterize the progress? What has surprised you?

Yann LeCun. I'm actually going to go back fifteen years to when Geoffrey Hinton, Yoshua Bengio, and I were trying to rekindle interest in neural networks, which we rebranded as "deep learning." And we were struck by how fast it was adopted: within eighteen months of the first papers showing that deep learning could improve speech recognition, basically every mobile speech recognition system was using a neural network.¹ This was my first exposure to how quickly technology is widely adopted when it actually brings something to the table.

There was a similar phenomenon in 2013 when convolutional neural networks became the rage for image recognition. Within months they were widely deployed, and people got excited about the prospect of driving assistance and autonomous driving. Within a few years, we had automatic emergency braking systems in cars, all using convolutional neural networks. These systems have become mandatory in Europe, leading to a 40 percent reduction in frontal collisions.² This technology was saving lives, and it made me proud.

The year 2015 was more about natural language processing (NLP). Yoshua's lab made people pay attention to attention mechanisms, which direct deep-learning models to focus their attention on the most relevant data.³ Then "Attention Is All You Need," which introduced the transformer architecture, surprised quite a few by building large networks as stacks of associative memory modules based on

so-called self-attention that compares inputs to each other.⁴ I had been advocating for self-supervised learning (SSL) for many years, but I did not expect that its success would be primarily in NLP – though we had hints that “shallow” SSL worked for text embedding, with Word2vec from Google and fastText at Meta.⁵ Then there was BERT from Google, which used transformers and attention instead of recurrent neural networks, which a lot of us suspected would not go very far.⁶

I wasn’t particularly interested in NLP myself, but it was certainly a very useful thing to do. What I found surprising was that a simple form of temporal prediction works so well for text. I had been working on temporal prediction for video for at least ten years before that, hoping that if you train a system to predict what’s going to happen in a video, it will understand how the world works: that the world is three dimensional and that objects move independently, for example. Some objects are inanimate and obey laws that are relatively easy to understand and make predictions about. Animated objects are much more difficult to predict. The hope and the whole idea of SSL is that by training a system through prediction, it can understand the underlying structure of what it is trained on. I was focusing on video because I believe that if we can do video, we can do everything. As it turns out, SSL from video is much harder than from text.

Manyika. As you think about where we are with the development of AI’s capabilities, are there scientific disciplines you’re most excited for AI to help make breakthrough advances?

LeCun. Certainly. Materials science, chemistry, biochemistry, and proteins are major topics. In medicine, can we understand and cure chronic diseases or cancers? Predicting protein structure is major progress, but the next step is designing proteins with particular properties.⁷ Can we build predictive mesoscopic AI models for complex systems like organelles, cells, or pathways in cells that are too complex for a single human to understand? The same applies to new materials and chemical compounds for improving energy storage (like batteries) or for catalysis (such as separating hydrogen from oxygen to store electrical energy as methane).

Importantly, the models that work well for proteins are really specialized; they’re not generic models. There is a hope that in the future we’ll come up with models that are more generic, that can make wider sets of predictions. This includes efficient prediction in large systems controlled by partial differential equations (PDEs) that are too complex for us to solve accurately and quickly. In fluid dynamics, for instance, sometimes you only need to predict lift and drag or turbulent behaviors. These collective behaviors can be captured by training a neural net (essentially 3D convolutional networks that mirror PDEs) to make predictions at a coarse scale. This allows simulating much larger or turbulent systems to predict general properties rather than requiring a detailed simulation.

Shirley Ho at the Flatiron Institute demonstrated a striking example of this.⁸ You can put the equations of general relativity into a simulator and simulate the first instants of the universe. Then you take that simulation and change the values of some of the parameters, like mass density or the properties of various fields. Then you run the simulation again to see if the result looks like our universe, if you get galaxies like those we see. That could validate or invalidate some theories about the nature of dark matter, for example. You can't do that simulation at the scale of the universe if you reduce it to the equations that we know. But if you run the simulation in a small cube, and you train a neural net to make a prediction of what goes on within that cube, but without requiring a detailed simulation of everything, then you can scale *that* up and perhaps run that simulation at a scale of the universe.

That is model prediction validation. The next step is design. If you have a predictive model, you can use inference by optimization to figure out the necessary interventions – action variables – to achieve a desired property for a chemical compound (for example, a specific energy separation of hydrogen from oxygen). This is the concept of a *causal world model* with very broad applications, such as in treating patients or controlling industrial processes and robots.

Manyika. What capabilities does AI need to help advance science and advance discovery?

LeCun. We can use machine learning to train predictive systems that can be phenomenological models. They're not necessarily models that explain *anything*, but they can make predictions that are sufficiently accurate and computationally much faster than a traditional reductionist model.

I think this is a crucial point and is what I am presently devoting all of my efforts to: devising AI systems that can find an abstract representation of the phenomenon and make predictions in that abstract representation space. This abstract representation eliminates a lot of details about the original observations. And that's a crucial point because LLMs (large language models) and other generative models are trained to predict *every detail* of the input. In language, it's not too much of a problem. You cannot predict exactly which word follows a sequence of words, but you can produce a probability distribution over words. That's easy because there's a finite number of possible words. But when you train the model to predict future frames in a video, you can't represent a useful distribution. You have to make predictions in an abstract representation space, not at the pixel level.

So a lot of people in the last few years instinctively said, "let's just tokenize the world." Let's take images from videos and cut them into little squares and turn that into a vector that doesn't look different from the one that represents a word, and feed this to a gigantic model to predict the next few frames. Frankly, it doesn't work that well. The reason why is that you simply cannot predict what's going to happen in a video at the pixel level. There are so many details that are just not in the

input. We don't know how to produce a probability distribution over all possible video frames because it's mathematically intractable. It's a problem people have struggled with for decades in statistical physics.

Instead, what we do as scientists is to find a representation of the input that eliminates all the details we cannot predict, and we make predictions in that representation space. That's not a generative architecture.

Manyika. Thinking about progress in AI, you've been vocal about the limitations you see in LLMs. What are those limitations, particularly in the context of applying AI to science?

LeCun. There are essentially three limitations. The first is that you cannot predict everything that takes place in the physical world or even in data collected from sensors. There is noise. Systems are stochastic – or we measure only partial information about the system. Trying to train a system to make an exact prediction in these conditions doesn't do anything particularly interesting. You have to let the system find the appropriate representation within which to make predictions. This is really what science is all about.

We could describe everything that takes place here between us in this conversation in terms of quantum field theory instead of, say, psychology. It would be completely impractical because we would need to measure the wave function of the universe within a cube that contains both of us. That's a pretty big thing. Then you'd need to run some insanely large quantum computations to make predictions. We never do this.

What we do is we invent abstractions that allow us to make predictions: particles, atoms, and then molecules; or proteins, organelles, cells, organs, organisms, societies, and ecosystems. You have this whole hierarchy. What defines a branch of science is the level of abstraction at which you work to describe reality.

The second limitation is that you want the system to be able to learn the equivalent of a mental model or its environment, or what many people call a "world model." Given your idea of the state of the world at time t and an action you imagine taking, a world model will predict the state of the world at time $t + 1$ resulting from that action. You can think of it as a causal model: state, action, next state.

If you have such a world model, you can use it for science. The theory is represented by this predictive model. I can perform a virtual experiment: If I put the world in a particular state, and take a sequence of interventions, what is the predicted result? Then I can do the experiment and verify in the real world what actually happens.

More important, if you have a world model, you can do reasoning and planning. You can specify a desired target state and use an optimization procedure to infer a sequence of interventions that will lead to that outcome state. We do this all the time

as humans, using our mental model of the world to plan a sequence of actions in a new situation that you've never faced before. This is what psychologists call System Two.⁹

With LLMs, you must train the system for all situations it may encounter, but a pure LLM doesn't really have the prediction and planning abilities I'm describing, of producing an output by optimization. There's no search in a pure LLM. Some, which are no longer pure LLMs, do reasoning through search by having the LLM produce lots of candidates for sequences of tokens, and then a separate neural net that rates them and picks the best one. A generate-and-select approach.

Manyika. In those cases, you use Monte Carlo tree search or more tournament-based systems in which you have multiple generations and search among them with a scoring function or other method of evaluating alternatives to then select among them. It's a form of search.

LeCun. That's right. You need two components. One neural net component predicts the most likely symbols that follow a sequence (actions or steps of reasoning). In games like chess and Go, you have a neural net that scores every possible move, and you only explore the ones that look the best. You need a component to generate candidate branches and a second component to evaluate them so that you can select the best ones and prune the exponentially growing tree of possibilities. The basic idea has been known since the 1960s, though training methods are key. Google has contributed a lot with AlphaGo and AlphaZero.¹⁰

For games, the objective is to win. You play a full game before you know if a sequence of moves was good. Reinforcement learning is all about how you attribute credit to individual moves going backward. If my value function at a given move ultimately proved not great, I need to propagate that outcome to tell the value function it was overestimating or underestimating. That's what reinforcement learning does.

In the context of LLMs, there's a similar but simpler process: the LLM produces lots of outputs and humans rate them – that's RLHF, or reinforcement learning from human feedback.¹¹ If the outputs are programs or mathematical proofs, you can run and check them. You can design the system to rate itself. But more often you need a human to say, "this one was good, this one was not good." Then the system adjusts parameters to give a higher score to the good outputs and a lower score for the bad ones.

This brings us to the third limitation. A lot of LLMs are post-trained using supervised learning, with experts providing answers to retrain the LLM. The problem is that as you increase the universe of questions, it gets ridiculously expensive. There's an infinite number of questions, and while you fine-tune for common ones, there's always a long tail of other questions and prompts not in your training set. Then you get nonsense. The system that appeared amazingly smart suddenly looks totally stupid. It appears not to know a super basic notion simply because it wasn't trained for a particular version of the question.

Manyika. You've been pursuing a different approach called JEPA (joint-embedding predictive architecture). Tell us more about it and why it's needed, especially for science.

LeCun. When training a generative architecture with self-supervised learning, you show the system a corrupted or transformed version of an input and train it to reconstruct the full input with all of its details. If you apply this to real data – which is multidimensional, continuous, and noisy, like video – you can't predict all the details. The trick is to learn an abstract representation that eliminates the details you cannot predict and perform predictions in that representation space.

Much of science is about choosing a good representation. Let me take a concrete example. Going back centuries, people were quite good at predicting the trajectory of planets and objects in the sky. It's a complex problem to determine a set of rules to describe how some planets appear to be going backward in the sky while others go forward. But if you change your representation of the problem and you know that Earth is just a planet rotating around the Sun, and the other planets also rotate around the Sun, it becomes obvious that some planets should sometimes go backward because you're going faster than them on your orbit. The representation is crucial. Another example is the ideal gas law, $PV = nRT$, which predicts the pressure of a gas as a function of temperature, assuming the molecules rarely bump into each other. This prediction is done in a representation space that eliminates almost everything about the underlying state of the system, like the position and velocities of all the gas molecules. We can't possibly measure them, so we ignore them. Everything we don't know about the system, we call entropy or heat. This whole idea – identifying what information is useful to make predictions and eliminating the stuff we cannot predict – is at the root of science and, I would argue, at the root of intelligence.

JEPA implements the idea of training a neural net to produce a representation of the input that may not contain all the details and simultaneously training a predictor in that representation space.¹² You make the system learn a representation that contains as much information as possible about the input but also allows you to make predictions: it does not contain all the stuff you cannot predict, like the motion of individual molecules in the gas example. We use an encoder to learn a representation space simultaneously with a predictor. The main point is: don't make predictions in input space, make predictions in representation space. The difficulty is preventing "collapse" in which the encoder ignores the input to produce a constant output, making prediction trivial.

Manyika. You used to work on energy-based models and entropy functions as ways to think about this. Is that still applicable here?

LeCun. The energy-based model view is the only way to interpret what goes on there. You can think of an autoencoder as an energy function that is the reconstruc-

tion error of an input. To feed it an input if the reconstruction error is zero or small means you've trained your autoencoder on this data. But if your reconstruction error is zero everywhere, that means it doesn't actually capture any interesting dependencies in the input. It just computes the identity function. You want an energy function that takes low energy around the data points that you train it on but takes *higher* values elsewhere. The energy becomes a contrast function that separates where your data are and where there are no data. Contrastive methods achieve this by generating points outside the manifold of data and pushing their energy up. Denoising autoencoders, like BERT-style NLP systems and masked autoencoders (MAE) for images/video, are examples. You take an input, corrupt it in some way to generate a point outside the manifold of data, and then you train your autoencoder to transform this corrupted version of your input into the original one.

An interesting version is diffusion models.¹³ When you corrupt the inputs, instead of training the autoencoder to directly reconstruct the original, you train it only to take you closer to the original. You start from a data point, then you noisify it. Now you tell a neural net: I want you to point this way so that when they give you this noisy point, you're making it less noisy. You do this for all the points in the trajectory so that you get a neural net to learn a vector field that takes you back to the data manifold. Or you could do this by training an energy function and then computing the gradient of this energy function, which will take you back to the data manifold as well.

Manyika. What you're describing is trying to learn and build representations from observations, which is intuitively appealing in the context of science. But in some areas of science, there may be sparse data or limitations with respect to observables. Does that limit AI's potential to advance science in those areas?

LeCun. The hope is you might be able to train a generic encoder and predictor, self-supervised, with lots of data from different sources, as long as the input representation does not vary widely. That's the power of pretraining: you have a generic model that is pretrained, and if you have data for a particular experiment, you fine tune it. You may not have huge data to train on, but the pretraining should put you in a good starting position.

Manyika. If you're training on data from different domains or at different levels of abstraction and doing reasoning at very different levels, how do you deal with the compositionality challenge?

LeCun. I think we would need hierarchical models. And when you go up a layer, you eliminate some details about the system's state representation at the level below. You have a less accurate representation, but that allows for longer-term predictions of more complex emergent phenomena. For the same reason, we don't use chemistry to explain human behavior. If we want to make long-term predic-

tions of complex systems, we need to abstract away details and find abstract representations. This is how you would train those systems: lower levels make short-term, detailed predictions. Higher levels are trained to make longer-term predictions for more complex systems. They eliminate more information to make those predictions. Deciding which level to use for each problem is a different story.

Manyika. One of the distinctions you've been making is between what you term AMI (advanced machine intelligence) and AGI (artificial general intelligence). Can you explain and contrast these concepts?

LeCun. First, I must say that there is no question in my mind that, at some point in the future, machines will be as intelligent as humans in all domains, including science.

But there is a question of vocabulary. I use the phrase *advanced machine intelligence* to mean the kind of intelligence that we observe in animals and humans. Not necessarily human-level, because even nonhuman animals are really smart. Current AI systems don't come close to that, at least in their understanding of the physical world and ability to plan.

Then there is human-level AI, which includes systems that have all the capabilities of humans, perhaps better in some domains. What people really mean by AGI is human-level AI, and there is a fallacy here: the hidden assumption that human intelligence is general. But human intelligence is not general at all. It's very specialized; it's what evolution found was useful for our species to survive. We know we don't have general intelligence. Otherwise, you wouldn't be able to buy a \$30 gadget that beats you at chess, right?

We're terrible at a lot of tasks that computer science has shown over the last seven or eight decades can be done by computers better than us. It used to be that you could have a career as a mathematician by essentially filling up logarithm tables. Now we have computers that do this much better than we can. It used to be common in science fiction novels, like Isaac Azimov's *I, Robot*, that the way you test the intelligence of a robot is to ask it to compute an integral symbolically.¹⁴ We can do this with computers now. It turns out that it's hard for humans, but it's not that hard computationally. I prefer to label advanced intelligence using the phrase artificial super intelligence (ASI) because it doesn't make any assumptions about the nature of human intelligence.

Manyika. Are you saying that AGI, as long as we compare that to human intelligence, is really a specialized intelligence as opposed to a general one – that it's really a subset of all intelligences?

LeCun. There can never be a completely general intelligence, for the same reason that in machine-learning theory, the “no free lunch” theorems say that there are only a small number of tasks that any learning system can learn efficiently.¹⁵ It's also

related to some more philosophical, conceptual things, like the fact that any measure of complexity you come up with will say that almost everything is random, except for a tiny number of things that your notion of complexity measures as simple.

That's at the root of thermodynamics. Only a small number of states of a physical system are considered organized. Just because of statistical fluctuations, it's going to go from what we consider organized to disorganized because there are many more disorganized states. Because of our notion of what is simple, only a small number of things are simple. I think there is a similar argument for intelligence. There cannot be a system that can solve every problem in the world. It has to be specialized. That's the philosophical aspect of it. So it's the word *general* that I just don't agree with.

Manyika. Terminology aside, you've critiqued the idea of getting to more capable intelligence all the way up to ASI via LLMs due to limitations of their autoregressive feed-forward paradigm. Can you explain that, particularly with respect to AI for science?

LeCun. The basic paradigm of intelligent inference that people in classical AI have known for decades is that you look for a solution to a problem by *search*. You need an intelligent machine with two components: one to check if a proposed output is good – *Is this a solution to my equation? Is it the shortest path between cities?* – and another to perform the search for the best solution. But there's no single magic bullet for every problem.

It took some time for the early pioneers of AI to realize this was a problem. In the 1950s, Alan Newell and Herbert Simon came up with this program that they very modestly called the General Problem Solver.¹⁶ It said that if I can formulate a problem in terms of a search, and I have a way of checking whether a proposed solution is a good one, then I can solve *any* problem. What they didn't realize is that the search problem is intractable most of the time. In the traveling salesman problem – in which you must find the shortest route through a group of cities, traveling through each only once and ending back at the city of origin – as you increase the number of cities, the complexity grows exponentially. That started the whole theoretical part of computer science, of complexity theory, when people realized that most problems that are interesting are actually intractable.

This idea that you produce an answer to a new problem by search is absolutely crucial, and LLMs don't do this. They produce autoregressively. There is no optimization procedure; there's no search in a search space. LLMs that produce lots of token sequences and rate which ones are best are using a very primitive form of this kind of search. That's the first essential thing. Second, this search should not be performed in token space or language, but instead in an abstract continuous representation space. That is the way humans operate and is relevant to AI in science.

Manyika. But to do search in representation space presupposes some kind of objective function. Yet for general intelligence or for science, we may not know what that is; we're often experimenting or simply exploring or following our curiosity. Does that suggest that because AI leaves out all that curiosity-driven work, there may be a limit to the use of these tools for science?

LeCun. I don't think it limits that at all, because what is curiosity? When you make a prediction about something that should happen as a consequence of an event or an action, there are two possibilities. If your prediction was wrong, your model is wrong; you need to adjust it and probably repeat the experiment.

But there is another thing that humans are capable of, and that I think future AI systems will be capable of, or should be capable of: predicting the reliability of a prediction. If there are certain actions for which your prediction of the outcome is uncertain, you should take the action, observe the result, and correct your model if your prediction was wrong. That is a form of curiosity. Infants learn about things like gravity around the age of eight months. Most six-month-olds haven't yet figured out that every object falls under gravity. But at ten months, infants have. If you put some eight-month-olds on highchairs with a bunch of toys, they will systematically throw the toys on the ground and watch them fall. They do the experiment and verify that gravity applies to everything. If you show them a helium balloon, a party balloon, and it floats, they are fascinated, because here is an object that violates the rule that they thought they learned.

That's where the notion of a world model comes in. There are two types: the unconditional and the conditional (or causal). In the unconditional type, you start with the state of the world at time t . The world is a dynamical system. Can I predict the state of the world at time $t + 1$? The longer the time horizon you want to run your prediction, the more difficult it is to make accurate, detailed predictions. What you can do is make predictions at a higher level of abstraction, where many details are not present. Then there are conditional models where, given the state of the world at time t and a hypothetical intervention, can I predict the state of the world at time $t + 1$? That's really what we do in experimental science. We have a system in a state, we do an experiment, and we look at the result. If our hypothesis is good, the prediction is accurate. If what our model predicts deviates from reality, our model needs adjustment. That's at the root of the process of science: finding a representation of reality that contains the relevant aspects of the system, doing an intervention, and predicting the next state. That's a causal model.

Manyika. The distinction you are suggesting here is that with the ability to probe, test, explore, or perturb the system, you get closer to building a causally derived sense of intelligence. Whereas one could argue that statistically derived intelligence tells you very little about the underlying nature and behavior of the system. Is that right?

LeCun. Yes. You can have observational predictive models that make predictions about how the world evolves, even without interventions. Astronomy is an example. It's not an experimental science; you can't actually influence the system you're observing. But you can still come up with models that are kind of causal. You can predict the trajectory of planets. The simplicity with which you can do this depends completely on the representation space that you choose. So predicting the trajectory of planets became a hell of a lot simpler once we figured out all the planets rotate around the Sun, including Earth. The crucial question of science is, can we come up with a good representation of reality that allows us to make predictions, and possibly predictions conditioned on actions? Any intelligent system needs to be able to do that.

Manyika. Given the advances in AI, and particularly if we go beyond human cognitive levels and AI systems come to understand more than we do, what are the implications for philosophy of science, how we do science, and the nature of scientific understanding?

LeCun. I think that question is not a new one. When we solved PDEs (partial differential equations) numerically with computers, did the computational fluid dynamics simulator understand physics better than we did? It can make a prediction and it's using an algorithm based on equations that humans came up with.

The next step AI enables is training a machine-learning system to make predictions from data without the manual step of reducing the process to equations. AI allows us to skip having to first build a model of reality that can then be computed. This is powerful because many phenomena in science are collective complex phenomena.

A pile of sand behaves in a particular way, and the theory for this is not entirely clear. The property of materials, particularly complex ones, cannot be directly derived from the elementary equations of quantum mechanics. It's just too complicated. Another example is the magic angle, 1.1 degrees, at which you rotate two stacked monolayers of carbon, called graphene, to form a superconductor. That's a collective phenomenon that is extremely difficult to explain. There are various properties of materials of this type that cannot be usefully reduced to a small number of equations from which you can derive this collective behavior. How does intelligence emerge from neurons in interaction? That's a philosophical question of how a super complex property like intelligence can emerge from a large number of relatively simple elements in interaction, but that's a pretty high-level thing. At a lower level are questions of how life emerges from the interaction between proteins. This transition is what has baffled scientists for a long time: the transition from the microscopic to the mesoscopic. This is where interesting things happen, like life, for example.

So now there's a new way of doing science, which is neither completely qualitative and observational nor reductionist, but is a data-driven, AI-powered phenomenological model that may allow us to bridge the gap between microscopic and macroscopic.

Manyika. In the future, do you imagine AI doing science and making discoveries by itself, or will it always be a tool for scientists to use for discovery?

LeCun. I think it's going to be a tool for a long time. There are probably some domains in which you can let the thing run for a long time and come up with something meaningful, and maybe it will be able to evaluate whether the observed behavior or the model is interesting. There could be some automation in some domains, but I think the relationship of scientists with AI systems for a long time will be the same as between a PhD advisor and a PhD student. Except that the goal of the latter relationship is for the PhD student to eventually be autonomous. "There is something interesting to explore in this area. Why don't you work on this and then come back when you get something interesting?" This is the difference between directing the research of a team and doing the research yourself.

AI will essentially operate as a staff of diligent research assistants. Even if these AI systems are better than us at solving individual problems or practicing scientific research – and it's great to work with people who are smarter than you! – we will still be their boss. We will still tell them what to work on, because ultimately, we use these tools to help ourselves as a species or society. Directing what AI systems produce is a role we'll preserve.

Manyika. In that world, could we be limited by the scientist who's conducting this orchestra of intelligent systems, whether they're PhD students or AI systems? Does that present some sort of ceiling to the science that we can do?

LeCun. No, I don't think that presents a ceiling, or not any more than an advisor is a ceiling to what a PhD student can produce. Sometimes PhD students come up with a cool idea the advisor hadn't thought about. You put them on the right track, but they redirect things a bit. We think about the entire process of science, so when individual scientists in an industry lab, for example, work on a topic, there is a little bit of influence from their managers, but not a huge amount, if it's a self-respecting research lab.

If there is too much influence on the scientists working in your lab, you are only going to get expected results. If you want breakthroughs, you have to give people a very long leash and basically get out of their way. That's a bit of the magic I observed at Bell Labs and tried to produce at FAIR (Meta's Fundamental AI Research): hire the best people, give them the means to succeed, orient them in a good direction, and get out of the way.

Manyika. Does that also imply that to progress AI and science, once they are capable enough, we should just "hire" the best AI systems, give them direction, and get out of the way?

LeCun. Only if you think we've already figured out all the main ideas and all we need to do is scale up and engineer the thing better. Frankly, a lot of people in AI are currently thinking that we already have the proper paradigm, and we just need to scale it up. I famously don't believe this. I believe that we need a few conceptual breakthroughs before we get to something like AMI or ASI. We need to do research that is not as corralled as the current default mode in the industry. Academia tends to be less directive, but there is still some direction and some constraints on the work you can do in that you need to raise money to pay for your students and postdocs, which generally comes from a particular program that NSF or some other agency decided was critical.

Manyika. You've been a very strong proponent of open science – for example, sharing preprints and open-source code. How does openness, in science and in AI, accelerate progress?

LeCun. The very process of science is centered around communication. You have an idea, you've done an experiment, you got some result, and you publish it. The reason you need to publish it is because you need other people in the community to be able to challenge the idea, try to verify the result themselves, and see if it's possible to do better. The progress of science relies on communication between scientists and therefore on publication.

If you do some groundbreaking research and you don't talk about it, it's as if you never thought about it. If you keep it a secret, *maybe* some years down the line you'll be able to build a product out of it. But how do you convince anyone that your idea is good and that they should fund you to realize your idea? You need to publish it. You need to explain: "Here is how it works and I tried it on this dataset, but I'm a poor PhD student in a university somewhere in a far corner of the world, and I can't test this on a big dataset without support." Here is the idea, someone else will have the means to bring it to something real, and then at some point it may spur some investment from industry and end up changing an industry or creating a new one. It's a long process, and people don't realize that process can take five, ten, twenty, thirty years. This was the case for deep learning.

By 2012, posting AI preprints on arXiv had become standard procedure; peer review and official publication would follow months later. The arXiv submission of the paper "Attention Is All You Need" had something like seven hundred citations before it was even presented at conference. I publicly claimed that if an advance is not published, it's not research, because you can fool yourself. I've seen this in industry research labs where ideas bubble up the hierarchy and appear amazing, but no one challenges or verifies them. Peer review is imperfect and feels brutal, but it can stop you from fooling yourself. The whole process of science is to prevent you from fooling yourself. It's really what the scientific method is all about.

The practice of open source has accelerated progress in AI in unimaginable ways. Tools like TensorFlow, PyTorch, and Jax have hugely enabled progress, and the fact that everyone publishes their code means you can reproduce and extend the work of others. That is behind the incredible speed of progress.

Manyika. For most scientists, the benefits are clear. But some in industry and government make arguments against open science based on competitiveness and intellectual property, as well as safety/geopolitical concerns. How do you respond?

LeCun. To the question of competition : There's a spectrum from blue sky (basic) research to engineering applications. Publications are more useful near the blue sky end of the spectrum. As you move toward application, the work requires know-how and details specific to your infrastructure. So there is a point at which if you want to keep your competitive edge, you don't want to publish. But you wouldn't publish anyway if you can't really communicate your application results through a classical paper.

The ability to publish also attracts the best scientists. You'll have a hard time hiring a self-respecting scientist if you forbid them from discussing their work. This is why famously secretive companies struggle to have top-level AI research groups. Also, the not-fooling-yourself aspect of science means higher-quality research generally. If scientists are mandated to publish, they will follow better methodology to survive peer review. And there's another consideration : the commercial impact of research might be several years down the line. If you want scientists to be motivated, they can't be motivated by product impact alone, because the probability of that happening can be small, even if theirs is a good idea. It could be twenty years down the line. Scientists have to score something sooner. Presenting a paper at a conference and earning citations are ways of scoring. And in industry, publications are a way for management to evaluate a piece of research. It is notoriously difficult to evaluate the potential long-term impact of a piece of research. Relying on the external research community can be a good way to avoid mistakes. If a paper from your lab wins an award at ICLR (International Conference on Learning Representations) or CVPR (Conference on Computer Vision and Pattern Recognition), you should probably pay attention to it internally. Same thing if an arXiv preprint earns two hundred citations in a few months ; or if that open-source package your product groups never paid attention to got ten thousand stars on GitHub and is used by the whole industry except you.

Now, about the defense and geopolitical question. On the spectrum from research to product or deployment, you could raise the threshold for making research public for geopolitical or safety reasons rather than for acceleration of scientific research. But you pay a price. The more secretive you become, the more you slow yourself down. And if you can do research and solve a problem in secret, then that problem probably was not that difficult to begin with. You might imagine having a monop-

only on good ideas to solve your problem, but if it's a hard problem, it will take the entire scientific community to take interest and exchange ideas.

The problem of ASI or AMI is of that nature. It's still a scientific question. It's not a question of scaling up or engineering resources. It's a scientific question, and nobody has a monopoly on good ideas, which is why I think some of the outfits that are claiming that they'll come up with AGI, or whatever they call it, within a few years, with a small team of ten people working in secret, that's complete BS. That just can't possibly happen. It's one of the most challenging scientific and technological problems of our times. You won't solve it with a handful of people working in secret; that only happens in bad sci-fi movies.

Manyika. As you look to the future, what are the big problems in science that you hope AI will have helped to solve?

LeCun. There are three major questions in science. What's the universe made of? What's life all about? And how does the brain work? Pretty much all of science falls into one of those three categories. I think the promise of AI for science is enormous.

ABOUT THE AUTHORS

Yann LeCun is Executive Chairman of Advanced Machine Intelligence (AMI) Labs and the Jacob T. Schwartz Professor of Computer Science, Data Science, Neural Science, and Electrical and Computer Engineering at New York University. He previously served as Chief AI Scientist for Meta's Fundamental AI Research (FAIR) team. He has recently published in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, *International Conference on Learning Representations 2025*, and *Advances in Neural Information Processing Systems*.

James Manyika, a Member of the American Academy since 2019, is SVP at Google-Alphabet and President for Research, Labs, Tech & Society.

ENDNOTES

- ¹ Geoffrey Hinton, Li Deng, Dong Yu, et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine* 29 (6) (2012): 82–97, <https://doi.org/10.1109/MSP.2012.2205597>.
- ² Brian Fildes, Michael Keall, Niels Bos, et al., "Effectiveness of Low Speed Autonomous Emergency Braking in Real-World Rear-End Crashes," *Accident Analysis & Prevention* 81 (2015): 24–29, <https://doi.org/10.1016/j.aap.2015.03.029>.

- ³ Dzmitry Bahdanau, Cho Kyunghyun, and Yoshua Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” arXiv (2014; rev. 2016), <https://arxiv.org/abs/1409.0473>.
- ⁴ Ashish Vaswani, Noam Shazeer, Niki Parmar, et al., “Attention Is All You Need,” arXiv (2017), <https://arxiv.org/abs/1706.03762>.
- ⁵ Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, “Efficient Estimation of Word Representations in Vector Space,” arXiv (2013), <https://arxiv.org/abs/1301.3781>; Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov, “Enriching Word Vectors with Subword Information,” arXiv (2016; rev. 2017), <https://arxiv.org/abs/1607.04606>; and Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov, “Bag of Tricks for Efficient Text Classification,” arXiv (2016; rev. 2017), <https://arxiv.org/abs/1607.01759>.
- ⁶ Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding,” arXiv (2018; rev. 2019), <https://arxiv.org/abs/1810.04805>.
- ⁷ John Jumper, Richard Evans, Katherine Tunyasuvunakool, et al., “Highly Accurate Protein Structure Prediction with AlphaFold,” *Nature* 596 (7873) (2021): 583–589, <https://doi.org/10.1038/s41586-021-03819-2>.
- ⁸ Siyu He, Yin Li, Yu Feng, et al., “Learning to Predict the Cosmological Structure Formation,” *Proceedings of the National Academy of Sciences* 116 (28) (2019): 13825–13832, <https://doi.org/10.1073/pnas.1821458116>.
- ⁹ Daniel Kahneman, *Thinking, Fast and Slow* (Farrar, Straus and Giroux, 2011).
- ¹⁰ David Silver, Aja Huang, Chris J. Maddison, et al., “Mastering the Game of Go with Deep Neural Networks and Tree Search,” *Nature* 529 (7587) (2016): 484–489, <https://doi.org/10.1038/nature16961>; and David Silver, Thomas Hubert, Julian Schrittwieser, et al., “A General Reinforcement Learning Algorithm that Masters Chess, Shogi, and Go through Self-Play,” *Science* 362 (6419) (2018): 1140–1144, <https://doi.org/10.1126/science.aar6404>.
- ¹¹ Long Ouyang, Jeffrey Wu, Xu Jiang, et al., “Training Language Models to Follow Instructions with Human Feedback,” arXiv (2022), <https://arxiv.org/abs/2203.02155>.
- ¹² Mahmoud Assran, Quentin Duval, Ishan Misra, et al., “Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture,” arXiv (2023), <https://arxiv.org/abs/2301.08243>.
- ¹³ Jonathan Ho, Ajay Jain, and Pieter Abbeel, “Denoising Diffusion Probabilistic Models,” arXiv (2020), <https://arxiv.org/abs/2006.11239>.
- ¹⁴ Isaac Asimov, *I, Robot* (Gnome Press, 1950).
- ¹⁵ David H. Wolpert and William G. Macready, “No Free Lunch Theorems for Optimization,” *IEEE Transactions on Evolutionary Computation* 1 (1) (1997): 67–82, <https://doi.org/10.1109/4235.585893>.
- ¹⁶ Allen Newell, John C. Shaw, and Herbert A. Simon, “Elements of a Theory of Human Problem Solving,” *Psychological Review* 65 (3) (1958): 151–166, <https://doi.org/10.1037/h0048495>.

From Alchemy to Alchemy: On Matter, Minds & Tools

Alán Aspuru-Guzik

This essay explores the impact of information processing on chemistry and materials science through a chronological journey that begins with the origins of life and ends with speculations of what will be possible in 2050. We pause at key moments in this history to reflect on changes to the three main, intertwined threads of this story: matter, minds, and tools. “Matter” refers to the chemical environment that our species interacts with. “Minds” refer to the collective of natural and artificial intelligence that we have access to, as well as our inherent biological information-processing capacity. “Tools” refer to the different processing technologies, such as flasks, oil refineries, computer processors, and so on, that humans employ to understand and manipulate matter. Through this framework, the essay considers the roles of cyborgs and androids today and speculates on the potential future of AI for science.

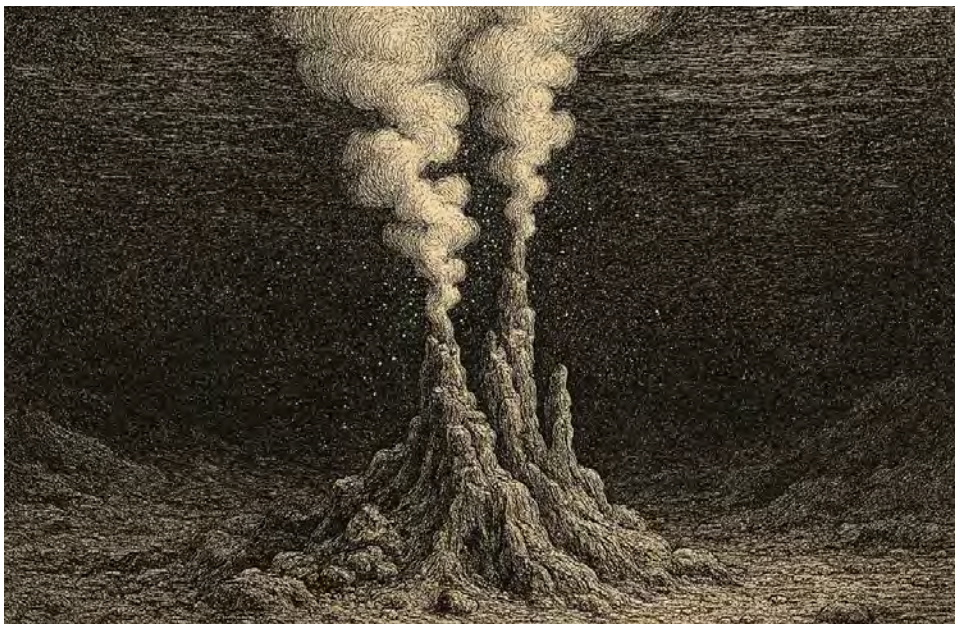
This story begins approximately four billion years ago, when life could have originated at an underwater hydrothermal vent, driven by chemical free-energy gradients generated by reactions between chemical compounds exiting the vent and those in the surrounding ocean (as seen in Figure 1).¹ This heterogeneity created an opportunity for the formation of steady-state chemical reaction networks composed of organic compounds, which, if persistent enough, could have led to the emergence of protolife. This protolife would eventually have learned to encapsulate information within membranes and developed an apparatus for self-replication.

This biological generative process highlights two critical elements of our discussion: *information*, which is related to the complexity of the chemical reaction network itself and the self-replication information that was achieved – this is when life began; and *matter*, which comprises the chemical substrates involved in metabolism and the making of the organism. Information and matter are the two essential components of life.

For most of those four billion years, information was mainly passed down through generations of different life forms by replication, a mechanism known as the fundamental dogma of biology. This process involves ribosomes that transcribe the information stored in deoxyribonucleic acid (DNA) into ribonucleic acid

Figure 1

Depiction of an Underwater Hydrothermal Vent



A depiction of a hydrothermal vent, where free-energy gradients could have originated life on Earth. Source: Image generated by the author using prompts in OpenAI's Sora.

(RNA) and then convert it into proteins. Proteins are responsible for the chemical networks that sustain cells, which eventually replicate DNA to produce progeny.

The natural “algorithm” that modifies this genetic information is called *evolution*. In this slow trial-and-error process, the probabilistic nature of replication generates a variety of organisms, some of which become more fit to survive while others become less so. The mutation rate of the genetic information reflects a trade-off between how fast an organism can try out new “solutions” to environmental problems and its own stability.² In addition, other forms of information are involved in life, for example, chemical communication between cells. These forms of information exchange were determined primarily by genetic information and therefore evolved at the same pace as evolution. Among bacteria, this phenomenon is known as quorum sensing.³

The impact of life on *matter* on Earth was tremendous. The origin of our oxygen-rich atmosphere dates back approximately two billion years to photosynthetic organisms. These early living organisms shaped the entire redox chemistry of the biosphere. To adapt to their environment, organisms grew complex nervous sys-

tems that allowed them to adapt to their environmental circumstances *on the fly*. Let us pause and consider the role of the *mind* in this context. Although the minds of dinosaurs and early mammals were quite sophisticated, and they communicated with each other, the crucial ability to process more complex information and pass it along to future generations was not possible, as far as we know, until the arrival of humans.

Fast-forward to two hundred thousand years ago: we can imagine a group of early hominids sitting around a fire and cooking food while experienced members discuss hunting techniques with younger ones, as shown in Figure 2 (left). Speech was believed to have evolved through a critical mutation in the FOXP2 gene, which may have been a pivotal step in the emergence of complex speech patterns.⁴

In our imagined scene, two key ideas in our discussion are represented: *information* and *matter*. Let us begin with *information*: Through *language*, knowledge and concepts gained from experience can be passed down to new generations in a completely novel way, much faster than through evolution alone. The accumulated information enabled by language facilitated interaction with the environment and other species, which led to rapid changes in prehistoric societies. Equally important in this scene are the chemical transformations that hominids began experimenting with: the use of fire enabled substantial changes in their diets, and the use of tools helped them further transform *matter*, for example, by aiding in the extraction of metals.⁵ Tool use, combined with language, therefore allowed our species to change its environment through increasingly complex chemical transformations.

The transformation of matter and the use of novel materials as tools are so essential that we define the ages of humanity by such tool-enabling materials. Here, we focus on the chemical transformations employed in conjunction with such novel materials. For example, Figure 2 (right) showcases a Mesoamerican woman making *masa* from nixtamalized corn. The process of nixtamalization involves soaking and cooking corn in *nextli* (from the Náhuatl word for volcanic or wood ashes, an alkaline source), which increases its nutritional value, improves its flavor, and enables its transformation into *tamalli*, or corn dough, for making tortillas.⁶ The *metate* shown in the figure is a stone-made tool that allows her to grind the nixtamalized corn and eventually flatten the *masa*. The *masa* is then cooked over fire to make tortillas. All these steps involve acid-base chemistry and fire to create one of the main staples of Mesoamerican civilizations.

Over the past two and a half million years, humanity has progressed from the Stone Age to the Bronze Age to the Iron Age and, within what is known as the Anthropocene, to the plastics age, the silicon age, and perhaps the era of materials by design.⁷ The written and archaeological records from these periods have revealed an intricate history of chemical transformations. For example, throughout history, humanity has utilized dyes to adorn clothing, decorate living

Figure 2

Depictions of Early Hominids and Humans Processing and Sharing Information



(Left) Hominids cook and talk around a fire. Two main themes of this essay are depicted in this image: first, the use of language to transfer information among generations; second, our use of fire to transform food and begin the era of human-induced chemical transformations. (Right) A Mesoamerican woman using a *metate* stone tool to prepare *masa* (corn dough). The *masa* had been previously chemically transformed by the nixtamalization process. These AI-generated images are meant to illustrate the intimate connection between cooking, chemistry, and our own development as a species. Source: Images generated by the author using prompts in OpenAI's Sora.

spaces and structures, and create a wide range of art forms. The ancient production of dyes and pigments originated from a complex set of biological and mineral sources.⁸ For instance, indigo is a plant-extracted dye used by several civilizations, from the Incas to the Egyptians.⁹ We now produce indigo synthetically, mostly to make our blue jeans blue.

For thousands of years, such chemical transformations were mainly empirical, and when they were not, they were described in terms of the philosophy of nature at the time. For example, to explain the origins of life, seventeenth-century Jesuit scholar and polymath Athanasius Kircher attributed them to a primordial “salino-sulphur-mercurial vapour,” which was the universal seed for not only inanimate but also animate objects.¹⁰ Half a century younger than Kircher, Isaac Newton was also an avid alchemist and devoted considerable time to his quest for the elixir of life and the philosopher's stone, which was believed to transform common metals into gold.¹¹

The transition into modern chemistry began around this time with scientists such as Robert Boyle and Robert Hooke, who were inspired by the emerging inductive philosophy of Francis Bacon, laying the groundwork for our current, systematic, and scientific approach to chemical problems.¹² One century later, in 1789, the use of induction and deduction led to the critical experiments of Henry Cavendish and Antoine Lavoisier, which dethroned the phlogiston theory and led to the discovery

of nitrogen and oxygen, respectively. The publication of Lavoisier's *Traité élémentaire de chimie* was as revolutionary as the French Revolution, which occurred the same year as its publication.¹³

We will skip the next couple of centuries. Although the field of chemistry has advanced substantially, including the development of chemical bond theories and the emergence of quantum chemistry, the way humans process information, by passing down their discoveries to new generations in verbal and written forms, has remained essentially unchanged until recently.¹⁴ The Industrial Revolution, in part, led to the creation of the bureaucratic state.¹⁵ The creation of the Computing-Tabulating-Recording-Company, now known as International Business Machines (IBM), was born out of the necessity of having an accurate population census. This, coupled with the need for extensive computing and tabulating operations during the Second World War, led to the birth of what we would now know as modern computing and eventually computer science.¹⁶ In what follows, we will intertwine the histories of chemistry and materials science with that of computer science. We will only briefly discuss the history of computing as it pertains to our story, focusing on its back-and-forth influence with chemistry and materials science. We will make stops at critical decades – the 1950s, the 1980s, the 2010s, and the current 2020s – and end the essay speculating about the intersection of these two fields in 2050.

During the twentieth century, the various subdisciplines, including organic, inorganic, and analytical chemistry, have established robust foundations and developed a rational framework for the understanding of chemicals and materials. For example, the pioneering work of chemist Christopher Ingold in the 1920s led to the concept of organic reaction mechanisms, which were designed to rationalize the complex behavior of matter. At Harvard University, organic chemist Robert Burns Woodward was a master at designing synthetic routes to increasingly complex natural products, accomplishing the synthesis of complex molecules such as cortisone, an immunosuppressant, in 1951, and reserpine, an antihypertensive, in 1956.¹⁷

The 1950s marked a pivotal period in the advancement of computational chemistry, as computers became increasingly accessible to scientists. In the same year that reserpine was synthesized, theoretical chemist Samuel Francis Boys and coworkers performed the first general quantum-chemical calculation on a modern computer.¹⁸ In 1957, physicists Berni Alder and Thomas Wainwright developed the first molecular dynamics code to study the physics of hard-sphere systems.¹⁹ The same year, Louis C. Ray and Russell A. Kirsch initiated the first cheminformatics database of chemical structures.²⁰ A couple of years later, theoretical chemists Christopher Longuet-Higgins and Lionel Salem investigated the alternating nature of chemical bonds in ethylene.²¹ These are just a few examples of the significant breakthroughs and publications that marked the beginning of the era of modern computer infor-

mation and chemistry (cheminformatics), as well as the interface of simulation and chemistry (computational and theoretical chemistry).

The Dartmouth Summer Research Project on Artificial Intelligence, a founding event in the field of AI, also began serious discussions in the 1950s to refine and advance ideas about thinking machines.²² At the time, deliberations were centered around cybernetics, a nascent field spearheaded by physiologist Arturo Rosenblueth and mathematician Norbert Wiener, whose concepts, such as feedback mechanisms and self-organization in complex systems, are explored in this issue of *Dædalus*.²³

Artificial intelligence and its subfields are highly relevant to the thread of this essay, as they represent human attempts to build machinery that can replicate human thought processes. This means that if the field is successful in generating powerful enough systems, information can now be processed, transformed, and communicated from AIs to other AIs as well as to humans. This would mark the beginning of the era of *androids* or artificial thinking beings. Even if partially successful, AI promises several societal changes. It is not surprising that, despite such great promise, the significant mismatch between the power of the computers available to researchers and researchers' algorithmic understanding of how to build intelligent systems has led the field to undergo several periods of excitement (springs) followed by disillusionment (winters).

This story now stops in the 1980s. With the advent of the microprocessor, computing became even more ubiquitous in science, thereby profoundly impacting chemistry. In the basement of Harvard's chemistry department, several scientists were working with organic chemist Elias James Corey on the LHASA (Logic and Heuristics Applied to Synthetic Analysis) computer program, a computational formalization of the concept of *retrosynthesis*: the idea that the (total) synthesis of a molecule can be broken down retroactively into the original fragments.²⁴ The computational "game" of retrosynthesis is not so dissimilar to the game of chess. This systematic integration of information and chemistry ultimately led to Corey receiving the 1990 Nobel Prize in Chemistry for his development of the theory and methodology of organic synthesis.²⁵

In the same decade, researchers made significant progress in the field of artificial intelligence. A former graduate student of Longuet-Higgins, computer scientist Geoffrey Hinton was studying Boltzmann machines and developing the powerful backpropagation approach that led to the modern theory of neural networks.²⁶ This work ultimately led to him receiving both the 2018 Turing Award (together with computer scientists Yoshua Bengio and Yann LeCun) for work on deep learning, and the 2024 Nobel Prize in Physics (with physicist John Hopfield) "for foundational discoveries and inventions that enable machine learning with artificial neural networks."²⁷ This Nobel Prize-winning work is an example of the formal connections between the theory of matter (statistical mechanics) and those of machine learning and artificial intelligence, which may be why in 1967 Longuet-Higgins switched

fields from theoretical chemistry to learning and cognition, where he made seminal contributions and also supervised Hinton. Unfortunately, although the decade of the 1980s saw huge progress, the results obtained during that period were not yet ready for prime time, as the data and computing capabilities were still not up to the requirements of the developed theory. This led to an AI winter that lasted for about two decades.

In the 2010s, another computational revolution began with the advent of the general-purpose graphics processing unit (GPGPU, or GPU for short) introduced by NVIDIA, which enabled scientists to perform more complex calculations on desktop computers. In 2012, at the University of Toronto, Hinton's team developed AlexNet, a neural network that became the state of the art for image recognition on the ImageNet dataset.²⁸ Around that time, there was a significant resurgence of interest in neural networks, including the hiring of AI researchers at top universities. For example, Harvard hired computer scientist Ryan P. Adams, a former postdoctoral researcher with Hinton, who cofounded the Whetlab machine learning startup while an assistant professor at the university from 2011 to 2016. I was also a faculty member at Harvard from 2006 to 2018, where I was lucky enough to participate in early research on AI for chemistry, as well as to engage with Longuet-Higgins's two career interests: *mind* and *matter*. Incidentally, I was trained as a computational chemist with an academic lineage that traced back to theoretical chemist John Pople, winner of a Nobel Prize in Chemistry and a collaborator of Longuet-Higgins.²⁹

After Adams arrived at Harvard, his interactions with students like David Duvenaud (a postdoc in his research group) and Rafael Gómez-Bombarelli (a postdoctoral fellow in my lab) – along with Adams's interactions with myself and many others outside of our groups – led to several key developments in the field of machine learning for chemistry. In particular, we developed a graph convolutional neural network for predicting molecular properties, which we successfully employed to discover organic light-emitting materials.³⁰ Our publication of this innovative network helped attract several machine learning researchers to the field of chemistry. Ten years later, it is common to see chemistry and materials science publications at the major computer science conferences such as the Conference on Neural Information Processing Systems (NeurIPS), the International Conference on Learning Representations (ICLR), and the International Conference on Machine Learning (ICML). Coincidentally, the room at Harvard where these publications originated as research, and which was also part of my research laboratory, is the same basement room that was once occupied by the students of Elias Corey, who developed the LHASA program forty years prior.

In 2016, we developed a generative model for the inverse design of molecules.³¹ Generative models are a potent and “magical” form of AI. Given a complex distribution, such as all the images or text on the internet or several voice recordings,

these models generate artificial samples from the distribution. In other words, the model can produce text, images, or voice recordings that did not exist in the training dataset yet are statistically similar to it. So far, it seems useful, but *not as useful* as when we can condition such a complex distribution on a desired set of attributes. For example, one could ask for audio samples that sound like the legendary singer Yma Sumac performing a song by José Alfredo Jiménez. Solving these types of problems is sometimes called *inverse design*, as one wants to create, from a distribution, novel samples that satisfy certain design conditions. This is in contrast to most computational methods, which employ *direct simulation* that links samples to their simulated properties. Inverse design opens the door to searching for (and finding) needles in a haystack.

Generative models for chemistry are extremely useful, given the enormous chemical space: It is, in principle, infinite, yet estimations of synthesizable, practical molecules are up to a range from 1,060 to 10,180. So the enumeration of these species is practically impossible using traditional methods. Here is where generative models shine.

Another key encounter during this part of my career at Harvard was with Alex Zhavoronkov, the CEO and cofounder of InSilico Medicine. At the time, Alex and his early-stage startup company were also experimenting with some of the earliest generative models in chemistry. Together, we established a collaboration that demonstrated the ability to generate a lead candidate for inhibiting the DDR1 kinase within forty-five days.³² This study and subsequent work laid the foundation for AI-based drug discovery. As of today, practically all the large pharma companies have significant AI drug-discovery programs.

Ten years ago, it was very common to say machine learning *tools*, as many of us didn't think of these models as *minds*. If we take the position that AI, even as of ten years ago, is a form of mind, it would still require physical tools to transform *matter*: even if it is possible for an AI system to suggest a chemical reaction for a human to carry out, it would be much more effective if the AI were empowered to carry out the reaction itself by means of automated equipment, analyze the results, and use its data for further refinement of a hypothesis or optimization of a desirable property. A significant number of scientists had already been thinking along these lines. In 2017, under the auspices of the global Mission Innovation initiative, which was spawned by the 2015 Paris Agreement on climate change, physicist and chemist Kristin Persson and I organized a conference in Mexico that led to a report that coined the term “materials acceleration platform,” now also known as a self-driving laboratory.³³ This conference included the participation of several pioneers of the field – such as Marty Burke, Jason Hein, Benji Maruyama, Tejs Vegge, and many others – and was the self-driving laboratory analogous moment to the original Dartmouth conference on artificial intelligence.

The concept of a self-driving lab represents a significant milestone. So far, the *human mind* has been the primary factor influencing the transformations of *matter*. However, as self-driving laboratories become increasingly intelligent and autonomous, these robotic platforms and devices have become an extension of humans, turning them into *cyborgs* – that is, humans augmented by machines – that conduct science.³⁴ Recent discoveries from self-driving labs include new conductive thin-film formulations, organic lasers, and perovskite nanocrystals.³⁵

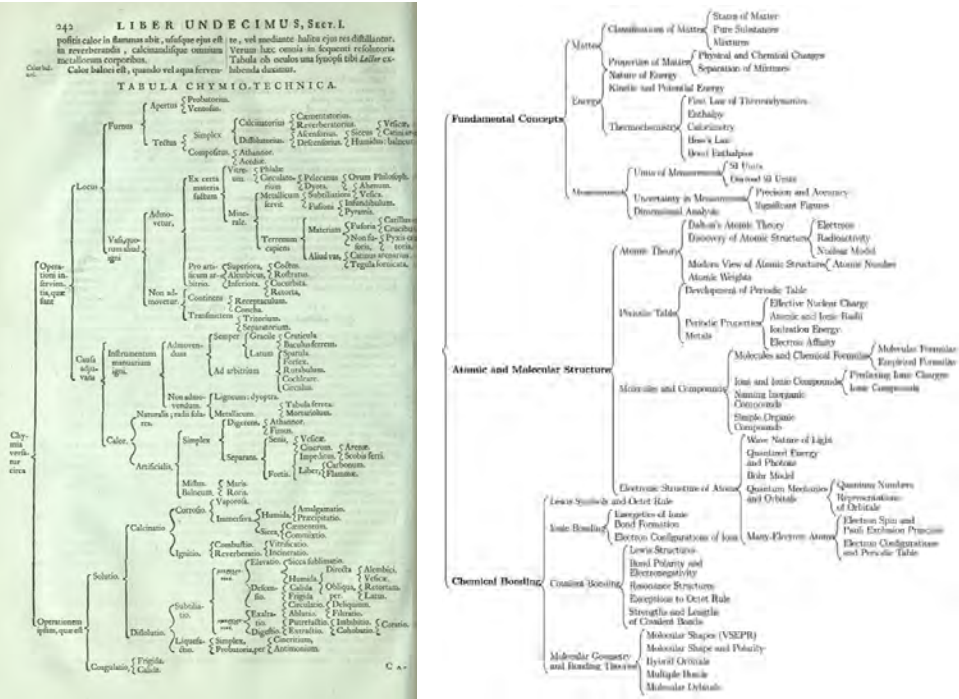
Back in Boston in 2018, inspired by the themes of the time, chemists Markus Reiher, Roland Lindh, and I published a manifesto outlining the new directions of research in computational chemistry for the twenty-first century, which we called “(R)evolution.”³⁶ In that work, we argue that theoretical chemistry in this century is closely tied to the technological revolutions of our time. We imagined a dialogue between a self-driving laboratory called Organa and a chemist named Jane. In this dialogue, the lab would interact with the chemist and help plan and execute an experiment. Confronting this challenge would require decades of community research. Nevertheless, just seven years later, Organa exists, and it performs experiments at our laboratory in Toronto (*vide infra*).

It is time to move forward to the current decade, the 2020s. This is the second issue of *Dædalus* this decade concerning AI, which is no coincidence given the current global focus on artificial intelligence and its implications for society.³⁷ Generative models of text and images have become ubiquitous and are predicted to transform almost every industry. The advent of large language models (LLMs) has transformed what is possible to achieve in AI at scale.³⁸ As an example of the application of LLMs in chemistry, let us compare summaries of the fields of alchemy and modern chemistry, as seen from the perspective of an undergraduate textbook. In Figure 3, we present (left) the classification of alchemy topics described by Athanasius Kircher in his 1665 scientific textbook *Mundus Subterraneus* and (right) an AI-generated summary tree built from a general chemistry textbook using TreeReader.³⁹

Now things come full circle. Athanasius Kircher spent a lifetime collecting what was known in the world for his “textbook” *Mundus Subterraneus*. He then further synthesized his knowledge into the tree-like table shown in Figure 3. Our tool TreeReader spent less than a minute summarizing a modern chemistry textbook. The striking side-by-side comparison of the differences and similarities, not only of how the trees were generated (human versus AI) but also how the field of chemistry has changed (alchemy versus the scientific method), refer deeply to the threads in this essay.

In a short time since the advent of LLMs, voice- and vision-controlled self-driving labs have become feasible. Based on these developments, and in collaboration with computer scientists Florian Shkurti and Animesh Garg, we have realized our dream

Figure 3
Alchemy and Modern Chemistry Compared Using Summary Trees



(Left) The classification of alchemical concepts as described by Athanasius Kircher in *Mundus Subterraneus, in XII libros digestus* (1665). (Right) An AI-derived summary of the classification of chemical concepts in a modern chemistry textbook. Source: Athanasius Kircher, *Mundus Subterraneus, in XII libros digestus* (Johannes Janssonius and Elizeus Weyerstraten, 1665), <https://dpul.princeton.edu/alchemy/catalog/2227mw62t>; and Zijian Zhang, Pan Chen, Fangshi Du, et al., “TreeReader: A Hierarchical Academic Paper Reader Powered by Language Models,” arXiv (2025), <https://doi.org/10.48550/arXiv.2507.18945>.

of developing and demonstrating Organa. We built Organa as an agentic assistant that can communicate with an end user to help plan electrochemistry experiments, surpassing our initial prediction of how long it would take to carry out this task!⁴⁰ Organa belongs to the large class of systems known as self-driving labs and is also part of a vast field of agents that conduct scientific research.⁴¹

A recent review highlights the vast array of agentic systems for science.⁴² Most of those agents are *minds* without physical *tools*. They utilize digital tools to conduct literature reviews, formulate hypotheses, and perform computational experiments. After developing Organa, we introduced El Agente, a computational chemistry agent

developed by over thirty people, as described by our graduate student Yunheng Zou and colleagues in “El Agente: An Autonomous Agent for Quantum Chemistry.”⁴³ As of today, El Agente can solve computational chemistry tasks at the level of an advanced undergraduate student. Connecting El Agente to self-driving labs is a current and meticulous work in progress, as one must carefully consider all possible safety guardrails.⁴⁴

Today, agents for science are operating in ways that are already transforming research, and they are on the verge of evolving into forms we can barely imagine. This is because large companies such as Anthropic, OpenAI, NVIDIA, and Google are continually adding base agentic features to their models, making them more capable of performing complex research tasks.

First, we can think of a straightforward *shallow* agent as a contextual description (which can be thought of as a “pre-prompt” or a general set of guidelines provided to a language model), followed by a set of instructions on how to approach a particular task or use a given tool. These agents could be valuable for mundane extract, transform, and load (ETL) operations, which remain among the most sought-after agent workflows, as data from academic groups and large companies can be convoluted in legacy documents that are not easily parsed. The notation of such papers can also be problematic. An example of a shallow agent we built in our Toronto lab is MERMaid, which extracts reaction conditions from legacy scientific PDF documents.⁴⁵ MERMaid builds a knowledge graph to store the extracted information, enabling it to construct a scientific world model across different publications.

For more complex tasks, such as generating hypotheses or focusing on the use of different tools, El Agente is considered a *deep* agent.⁴⁶ By that, I mean it is built on a hierarchical cognitive architecture inspired by frameworks like Soar and CoALA, designed to provide LLM agents with structured reasoning and long-term autonomy.⁴⁷ At its core are several interlinked types of memory. The working memory combines four elements: 1) a global memory that stores shared user directives and molecular details, ensuring all agents maintain a coherent context; 2) agent-specific conversation histories that record local decisions and dialogues; 3) a grounding mechanism that links agents’ reasoning to the actual computational environment (such as file structures); and 4) a long-term memory divided into procedural, semantic, and episodic layers. Procedural memory encodes how agents collaborate to perform tasks, semantic memory stores patterns and domain knowledge, and episodic memory retains past experiences to facilitate continual learning. Together, these memory systems enable the architecture to function in a manner that mirrors cognitive processes – planning, reasoning, and self-critique – across multiple timescales.

This architecture functions as a hierarchy of agents working in concert. At the top sits the computational chemistry agent, responsible for high-level planning and interpreting the user’s natural-language goals. Beneath it, mid-level agents

manage specialized domains – such as geometry generation and optimization, quantum calculations, and file management – analogous to departments in a complex organization. Each of these mid-level agents oversees lower-level expert agents that execute concrete actions, such as generating quantum chemistry input files, submitting jobs to high-performance computing clusters, or diagnosing errors. Communication flows downward through task delegation and upward through feedback and summarized reporting, allowing the system to maintain clarity even in multistep workflows. This cybernetic network design strikes a balance between autonomy and coordination: each agent acts locally, using its own contextual memory and reasoning core, yet collectively they form a coherent, adaptive whole. The result is a system capable of long-term reasoning, adaptive error correction, and transparent action logging, making the scientific process both traceable and intelligible to human collaborators.

Let's pause for a moment to further discuss the tools that El Agente employs. These quantum chemistry and molecular simulation tools are approximate *digital twins* of the quantum world. They solve the Schrödinger equation numerically, using an approximate method. The process of employing these programs is quite intricate. Therefore, a deep agentic system is needed to describe the input and parse the output of these programs, which comprise up to millions of lines of numerical code. Incidentally, these codes are the descendants of the computer packages that Samuel Boys developed in the 1950s.⁴⁸ For a medium-sized molecule, these calculations can consume hours of computer time on a modern computer cluster. Therefore, agents for science can end up orchestrating complex and costly computational workflows.

To make El Agente more closely resemble a scientist, we need to add hypothesis-generation capabilities to the tool. An ongoing project that bridges this gap is called PaperMon, a dual homage to the demons of science – imagined beings summoned in thought experiments to challenge scientific theories, such as Maxwell's demon – and Pokémon.⁴⁹ PaperMon enables the instantiation of a small demon per scientific paper. These agents have contextual information about a particular scientific paper or set of documents and become advocates for them. We then built a system that enables PaperMon to discuss with each other and generate sensible hypotheses. There is a significant gap between what we do nowadays with these tools and the work of a full-blown AI scientist. Still, the path between them is clear: the artificial scientist needs to be able to *generalize* to different problems and to clearly and understandably describe the scientific concepts it understands to others.⁵⁰

During the months I have spent writing this essay, large-scale advances in venture capital investment in AI and materials science have sparked renewed global interest in their intersection. Companies like Lila Sciences, Periodic Labs, Orbital Materials, and CuspAI have received more investment than the entire annual budget of the U.S. National Science Foundation in the fields of

chemistry and materials science.⁵¹ The United States has also announced the Genesis Mission, a project being compared in scope to the Manhattan Project, to advance AI for science, with materials as a primary focus.⁵² The year 2026 is when all these threads come together, attracting both private and public interest.

Making predictions is a risky business, but I would like to consider what the intersection of artificial intelligence and chemistry will look like in 2050. We may have already developed AI systems that match or surpass a human's ability to generate hypotheses and design experiments. Furthermore, the embodiment of these AIs would enable them to conduct the experiments themselves, thereby creating fully autonomous, self-driving laboratories. These systems would not be *cyborgs* or human-machine hybrids but rather full-blown *androids* or AI scientists.

As we conclude our chronological journey, it is worthwhile to revisit the geophysical impact of the evolution of oxygen-evolving organisms. They entirely transformed our biosphere. It is also worthwhile to speculate and reflect on potential implications for our environment from these large-scale AI scientists. Data center energy consumption is expected to account for a significant fraction of global energy demand by 2050. Suppose AI scientists succeed in undertaking noble scientific and engineering endeavors. In that case, they will have contributed to humanity's transition to a green economy by discovering new materials and engineering them into sustainable energy systems that align with the survival of our ecosystems.

On the other hand, if the current "business as usual" trajectory holds, we may need to devise a new term for the era that could follow the Anthropocene. For example, the *Alocene* would be an era in which AI itself is responsible for large-scale changes to the biosphere, potentially and irreversibly altering ecosystems. Here, what started as a computational *tool* would have become a *mind* of its own, capable of using its own tools (self-driving labs) and affecting *matter* at a planetary scale through its own energy consumption. And this is how our story would come full circle, once again.

AUTHOR'S NOTE

Most of the images in this essay were generated as a *cyborg* between simple prompts and OpenAI's text-to-video model Sora. They are inspired by book illustrations from the 1800s, which in turn take me back to the readings of several Jules Verne books I devoured cover to cover when I was around ten years old. I dedicate this work to Varinia Bernales for the valuable discussions that inspired several aspects of this manuscript, as well as her support and edits that helped bring it to completion. I would also like to congratulate Zijian Zhang for generating Figure 3 (right) using his tool TreeReader. Finally, I would like to express my gratitude to Dr. Anders G. Frøseth for his continued support of my scholarship and his friendship.

ABOUT THE AUTHOR

Alán Aspuru-Guzik is Director of the Acceleration Consortium and Professor of Chemistry and Computer Science at the University of Toronto. He is also a Senior Director of Quantum Chemistry at NVIDIA. He has published in such journals as *Science*, *Nature*, and *Proceedings of the National Academy of Sciences*.

ENDNOTES

- ¹ William Martin, John Baross, Deborah Kelley, and Michael J. Russell, "Hydrothermal Vents and the Origin of Life," *Nature Reviews Microbiology* 6 (11) (2008): 805–814.
- ² Dalit Engelhardt and Eugene I. Shakhnovich, "Mutation Rate Variability as a Driving Force in Adaptive Evolution," *Physical Review E* 99 (2019): 022424.
- ³ Michiko E. Taga and Bonnie L. Bassler, "Chemical Communication Among Bacteria," *Proceedings of the National Academy of Sciences* 100 (S2) (2003): 14549–14554.
- ⁴ Michael P. Richards, "A Brief Review of the Archaeological Evidence for Palaeolithic and Neolithic Subsistence," *European Journal of Clinical Nutrition* 56 (2002): 1262–1278; and Rafael Mier, *Maíz: Origen, cultura y cocina* (Ediciones Larousse, 2025).
- ⁵ Richards, "A Brief Review of the Archaeological Evidence for Palaeolithic and Neolithic Subsistence."
- ⁶ Mier, *Maíz*.
- ⁷ Colin N. Waters, Jan Zalasiewicz, Colin Summerhayes, et al., "The Anthropocene Is Functionally and Stratigraphically Distinct from the Holocene," *Science* 351 (6269) (2016), <https://doi.org/10.1126/science.aad2622>; Jennifer A. Brandon, William Jones, and Mark D. Ohman, "Multidecadal Increase in Plastic Particles in Coastal Ocean Sediments," *Science Advances* 5 (9) (2019), <https://doi.org/10.1126/sciadv.aax0587>; and Troy Scott, Amanda Walsh, Benjamin Anderson, et al., "High-Tech Infrastructure and Economic Growth: The Materials Genome Initiative," *Science and Public Policy* 48 (5) (2021): 649–661.
- ⁸ Omar Abdel-Kareem, "History of Dyes Used in Different Historical Periods of Egypt," *Research Journal of Textile and Apparel* 16 (4) (2012): 79–92.
- ⁹ Jeffrey C. Splitstoser, Tom D. Dillehay, Jan Wouters, and Ana Claro, "Early Pre-Hispanic Use of Indigo Blue in Peru," *Science Advances* 2 (9) (2016), <https://doi.org/10.1126/sciadv.1501623>.
- ¹⁰ Daniel Stolzenberg, ed., *The Great Art of Knowing: The Baroque Encyclopedia of Athanasius Kircher* (Stanford University Libraries, 2001); and Hiro Hirari, "Kircher's Chymical Interpretation of the Creation and Spontaneous Generation," in *Chymists and Chymistry: Studies in the History of Alchemy and Early Modern Chemistry*, ed. Lawrence M. Principe (Science History Publications, 2007), 77–87, <https://backoffice.biblio.ugent.be/download/376797/6797072>.
- ¹¹ Richard S. Westfall, *Never at Rest: A Biography of Isaac Newton* (Cambridge University Press, 1983).
- ¹² John Read, *From Alchemy to Chemistry* (Dover Publications, 2003).
- ¹³ Antoine Lavoisier, *Traité élémentaire de chimie* (1789).
- ¹⁴ Kostas Gavroglu and Ana Simões, *Neither Physics nor Chemistry: A History of Quantum Chemistry* (MIT Press, 2011).

- ¹⁵ David Graeber, *The Utopia of Rules: On Technology, Stupidity, and the Secret Joys of Bureaucracy* (Melville House, 2015).
- ¹⁶ Peter Galison, *Image and Logic: A Material Culture of Microphysics* (University of Chicago Press, 1997).
- ¹⁷ Ibid. ; and R. B. Woodward, F. E. Bader, H. Bickel, et al., “The Total Synthesis of Reserpine,” *Journal of the American Chemical Society* 78 (9) (1956): 2023–2025.
- ¹⁸ Woodward, Bader, Bickel, et al., “The Total Synthesis of Reserpine.”
- ¹⁹ Berni J. Alder and Thomas E. Wainwright, “Phase Transition for a Hard Sphere System,” *The Journal of Chemical Physics* 27 (5) (1957): 1208–1209.
- ²⁰ Louis C. Ray and Russell A. Kirsch, “Finding Chemical Records by Digital Computers,” *Science* 126 (3278) (1957): 814–819.
- ²¹ Hugh Christopher Longuet-Higgins and Lionel Salem, “The Alternation of Bond Lengths in Long Conjugated Chain Molecules,” *Proceedings of the Royal Society A* 251 (1265) (1959): 172–185.
- ²² Ronald Kline, “Cybernetics, Automata Studies, and the Dartmouth Conference on Artificial Intelligence,” *IEEE Annals of the History of Computing* 33 (4) (2011): 5–16.
- ²³ Arturo Rosenblueth, Norbert Wiener, and Julian Bigelow, “Behavior, Purpose and Teleology,” *Philosophy of Science* 10 (1) (1943): 18–24.
- ²⁴ Elias J. Corey, Alan K. Long, and Stewart D. Rubenstein, “Computer-Assisted Analysis in Organic Synthesis,” *Science* 228 (4698) (1985): 408–418.
- ²⁵ The Nobel Foundation, “Nobel Prize in Chemistry 1990,” <https://www.nobelprize.org/prizes/chemistry/1990/summary> (accessed February 17, 2026).
- ²⁶ David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski, “A Learning Algorithm for Boltzmann Machines,” *Cognitive Science* 9 (1) (1985): 147–169; and Werner Kutzelnigg, “H. C. Longuet-Higgins and J. A. Pople: Theoretical Chemistry,” *Angewandte Chemie International Edition* 43 (21) (2004): 2740–2743.
- ²⁷ Association for Computing Machinery, “Fathers of the Deep Learning Revolution Receive ACM A. M. Turing Award,” <https://awards.acm.org/about/2018-turing> (accessed February 17, 2026); and The Nobel Foundation, “The Nobel Prize in Physics 2024,” <https://www.nobelprize.org/prizes/physics/2024/press-release> (accessed February 17, 2026).
- ²⁸ David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, et al., “Convolutional Networks on Graphs for Learning Molecular Fingerprints,” in *Advances in Neural Information Processing Systems* 28 (NIPS 2015), ed. Corinna Cortes, Neil Lawrence, Daniel Lee, et al. (Curran Associates, Inc., 2015), <https://www.proceedings.com/content/030/03077owebtoc.pdf>.
- ²⁹ Kutzelnigg, “H. C. Longuet-Higgins and J. A. Pople”; and The Nobel Foundation, “John Pople–Facts,” <https://www.nobelprize.org/prizes/chemistry/1998/pople/facts> (accessed February 17, 2026).
- ³⁰ Duvenaud, Maclaurin, Aguilera-Iparraguirre, et al., “Convolutional Networks on Graphs for Learning Molecular Fingerprints”; and Alex Zhavoronkov, Yan A. Ivanenkov, Alex Aliper, et al., “Deep Learning Enables Rapid Identification of Potent DDR1 Kinase Inhibitors,” *Nature Biotechnology* 37 (9) (2019): 1038–1040.

- ³¹ Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, et al., “Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules,” *ACS Central Science* 4 (2) (2018): 268–276; and Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik, “Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering,” *Science* 361 (6400) (2018): 360–365.
- ³² Zhavoronkov, Ivanenkov, Aliper, et al., “Deep Learning Enables Rapid Identification of Potent DDR1 Kinase Inhibitors.”
- ³³ Mission Innovation, “Press Release—Materials Workshop Report,” January 26, 2018, <https://mission-innovation.net/press-release-materials-workshop-report>; and Gary Tom, Stefan P. Schmid, Sterling G. Baird, et al., “Self-Driving Laboratories for Chemistry and Materials Science,” *Chemical Reviews* 124 (16) (2024): 9633–9732.
- ³⁴ Benjamin P. MacLeod, Fraser G. L. Parlane, Thomas D. Morrissey, et al., “Self-Driving Laboratory for Accelerated Discovery of Thin-Film Materials,” *Science Advances* 6 (20) (2020), <https://doi.org/10.1126/sciadv.aaz8867>.
- ³⁵ Ibid.; Felix Strieth-Kalthoff, Han Hao, Vandana Rathore, et al., “Delocalized, Asynchronous, Closed-Loop Discovery of Organic Laser Emitters,” *Science* 384 (6697) (2024), <https://doi.org/10.1126/science.adk9227>; and Sina Sadeghi, Karl Mattsson, Joshua Glasheen, et al., “A Self-Driving Fluidic Lab for Data-Driven Synthesis of Lead-Free Perovskite Nanocrystals,” *Digital Discovery* 4 (7) (2025): 1722–1733.
- ³⁶ Alán Aspuru-Guzik, Roland Lindh, and Markus Reiher, “The Matter Simulation (R)evolution,” *ACS Central Science* 4 (2) (2018): 144–152.
- ³⁷ James Manyika, “Getting AI Right: Introductory Notes on AI & Society,” *Dædalus* 151 (2) (Spring 2022): 5–27, <https://www.amacad.org/publication/daedalus/getting-ai-right-introductory-notes-ai-society>.
- ³⁸ Ashish Vaswani, Noam Shazeer, Niki Parmar, et al., “Attention Is All You Need,” in *Advances in Neural Information Processing Systems* 30 (NIPS 2017), ed. Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, et al. (Curran Associates, Inc., 2017).
- ³⁹ Jiaqi Wei, Yuejin Yang, Xiang Zhang, et al., “From AI for Science to Agentic Science: A Survey on Autonomous Scientific Discovery,” arXiv (2025), <http://arxiv.org/abs/2508.14111>; Yunheng Zou, Austin H. Cheng, Abdulrahman Aldossary, et al., “El Agente: An Autonomous Agent for Quantum Chemistry,” *Matter* 8 (7) (2025), <https://doi.org/10.1016/j.matt.2025.102263>; Athanasius Kircher, *Mundus Subterraneus, in XII libros digestus* (Johannes Janssonius and Elizeus Weyerstraten, 1665), <https://dpul.princeton.edu/alchemy/catalog/2227mw62t>; and Theodore E. Brown, H. Eugene LeMay, Bruce E. Bursten, et al., *Chemistry: The Central Science*, 12th ed. (Pearson Education, 2013).
- ⁴⁰ Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas Griffiths, “Cognitive Architectures for Language Agents,” *Transactions on Machine Learning Research* (2024), <https://openreview.net/pdf?id=1i6ZCvflQJ>.
- ⁴¹ Tom, Schmid, Baird, et al., “Self-Driving Laboratories for Chemistry and Materials Science.”
- ⁴² Wei, Yang, Zhang, et al., “From AI for Science to Agentic Science.”
- ⁴³ Zou, Cheng, Aldossary, et al., “El Agente.”
- ⁴⁴ Sumers, Yao, Narasimhan, and Griffiths, “Cognitive Architectures for Language Agents.”
- ⁴⁵ John E. Laird, *The Soar Cognitive Architecture* (MIT Press, 2021).
- ⁴⁶ Zou, Cheng, Aldossary, et al., “El Agente.”

- ⁴⁷ Laird, *The Soar Cognitive Architecture*; and Sumers, Yao, Narasimhan, and Griffiths, “Cognitive Architectures for Language Agents.”
- ⁴⁸ Samuel F. Boys, Gregory B. Cook, Colin M. Reeves, and Isaiah Shavitt, “Automatic Fundamental Calculations of Molecular Structure,” *Nature* 178 (4544) (1956): 1207–1209.
- ⁴⁹ Mario Krenn, Robert Pollice, Si Yue Guo, et al., “On Scientific Understanding with Artificial Intelligence,” *Nature Reviews Physics* 4 (12) (2022): 761–769.
- ⁵⁰ Ibid.
- ⁵¹ Charles Yang, “Venture Capital is Subsidizing U.S. Material Science Research,” ML4Sci, October 15, 2025, <https://ml4sci.substack.com/p/venture-capital-is-subsidizing-us>.
- ⁵² The White House, “Launching the Genesis Mission,” November 24, 2025, <https://www.whitehouse.gov/presidential-actions/2025/11/launching-the-genesis-mission>.

Unlocking Scientific Intuition & Reasoning at Digital Speed

Pushmeet Kohli

Throughout history, scientific progress has been defined by the tools we build to extend our perception. Today, artificial intelligence stands as our newest instrument, not merely as a computational engine but as a revolutionary lens for deciphering the nature of reality. Drawing on my research journey at Google DeepMind, this essay explores how scientists are deploying AI to overcome the hardest “root node” bottlenecks in science. By applying machine intuition, we are uncovering breakthroughs across biology, materials science, and mathematics, from solving the grand challenge of protein folding with AlphaFold to charting new universes of stable materials with GNoME. The essay concludes by reflecting on the transition toward open-ended, agentic AI systems that actively generate novel hypotheses and their implications for the nature of scientific inquiry. We are evolving from the solvers of intricate puzzles into the architects of profound questions, embarking on the most exciting journey of discovery ever undertaken.

Humanity has come far. In a blink of planetary time, we have filled libraries, split atoms, built computing machines, and mapped genomes. On the scaffold of hard-won scientific insight we have constructed a thriving, global civilization. The mechanism that made all this progress possible – the human brain, with its astonishing abilities of perception, intuition, and reasoning – has enabled us to understand many of the processes and laws of our universe and use them to transform the world we live in.

Despite these advances, the challenges facing humanity today – climate change, pandemics, the need for sustainable energy – show that we need a deeper understanding of the world than we have ever had. The quest to understand the nature of reality is challenging the limits of our intuition and our reasoning abilities.

Throughout history, our tools have helped us transform civilization by extending what we could perceive, understand, and control. The telescope revealed the vastness of the cosmos; the microscope, the hidden architectures of life. The computer, the bedrock of modern society, made it possible to calculate, simulate, and accumulate a universe of scientific data whose vastness and subtlety can be overwhelming.

Artificial intelligence is the latest instrument in this lineage: a new lens on our perception of reality itself; a telescope that not only is powerful enough to scan the

vastness of the data-universe but can focus on what matters within that universe, delivering new levels of understanding and transforming research in areas like biology, materials science, and mathematics.

AI naturally dovetails with science because many of the hardest scientific problems are, at their core, search and reasoning problems. When we engage in scientific inquiry, we observe the world and generate and test hypotheses, refining or discarding them based on how well they explain what we see. But as the volume and complexity of scientific data have increased, this process has hit two roadblocks: 1) the scale of what we are trying to understand and 2) the speed at which we can test our ideas. AI overcomes these barriers because it not only can process an extraordinary volume of observations about the world but can propose and evaluate novel hypotheses at digital speeds.

Where traditional methods of inquiry relied on human ingenuity, often guided by sparse data, AI can now detect structure in overwhelming abundance. What emerges is something akin to *machine intuition*: a powerful, learned strategy for navigating immense complexity.

Because it is domain agnostic, AI also bridges disciplines. Unlike human experts, whose training is often deep but narrow, AI systems can ingest and integrate information across fields. Polymaths are rare; experts in both genomics and high-energy physics are virtually nonexistent. But because AI can draw on insights from both areas at once, it has a unique ability to make new connections. This, in turn, accelerates the interdisciplinary breakthroughs on which modern science increasingly depends.

A senior scientist recently confided to me, “I was skeptical of AI, but having experienced the speed and depth of insights AI tools bring, I find myself using them extensively in my work.” That admission brought to mind how early rocket scientists must have felt when they first got access to digital computers. That leap in capability – from extensive manual trajectory calculations using ranks of human “computers” to digital computation – must have been as empowering as it was irresistible.

This is not to say that the current generation of AI systems is a panacea or the sudden solution to all the world’s problems. They remain a tool in the hands of scientists; a powerful tool, to be certain, but a tool nonetheless.

For over two decades, I have immersed myself in machine learning and artificial intelligence. My research career began with a focus on perception and reasoning, endowing computers with the ability to “see” the world and building self-correcting systems that ensure software reliability. But while I found the mathematical and technical puzzles exciting, I harbored a lingering desire to understand the deeper nature of our world.

My research journey changed during a chance encounter with Demis Hassabis, the cofounder and CEO of DeepMind. When he shared his vision of AI not just as

a computational tool but as the ultimate lens for deciphering the nature of reality, I joined the mission.

DeepMind had already demonstrated the power of deep reinforcement learning: first by mastering Atari 2600 games and then through AlphaGo. By defeating one of the world's greatest Go players with moves that defied centuries of human wisdom – such as the now-famous “move 37” – AlphaGo proved that AI could transcend human intuition.

Believing the technology was finally ready to tackle the hardest challenges in science, Demis encouraged me to establish and lead the AI for Science program at Google DeepMind. Drawing on my experience with AI for Science, this essay traces the trajectory of our recent breakthroughs and the principles that have guided them. It offers a field guide to how AI currently integrates with the scientific method – where it excels and where it must improve – and explores how the emergence of agentic systems will fundamentally reshape the future of discovery. The following sections share my reflections on this new era of scientific discovery through deep dives into three key fields – biology, materials science, and mathematics – before zooming out to consider how AI is reshaping the scientific process itself, where its current limits lie, and what the future holds.

Biology is a natural starting point to test the utility of AI. It is staggeringly complex and full of emergent patterns and hidden structures that challenge the limits of perception and intuition. From understanding proteins – the building blocks of life – to understanding the genome – the code of life – biology has challenged us with profound mysteries. Among these, one of the most challenging was the fifty-year “grand challenge” of predicting the structure of proteins.

Proteins are the molecular machines of life. Made of intricately folded chains of amino acids, they are responsible for nearly every biological function, from digesting food to fighting infection. To understand disease, design better drugs, and ultimately engineer the machinery of life itself, we must know how proteins fold.

While a protein's amino acid sequence is easy to determine from its genetic code, its folded 3D shape is another matter entirely: each chain can potentially twist and bend in an astronomical number of ways, requiring analysis beyond the capabilities of any classical computation system. Discovering the structure of even a single protein through lab work can take years of effort using specialized equipment, at a cost of hundreds of thousands of dollars. For decades, scientists had tried and failed to accurately predict a protein's 3D structure directly from its DNA sequence using computation alone.

As a problem for AI, protein folding has all the right ingredients: a vast search space of potential structures, a clear and measurable goal (atomic-level accuracy), and a rich supply of protein structure training data gathered through decades of painstaking experimental work.

Over several years, my team at DeepMind worked closely with structural biologists and domain experts, including collaborators at the Protein Structure Prediction Center, whose Critical Assessment of Protein Structure Prediction (CASP) experiment set the global benchmark for protein prediction. In 2018, our first AI model, AlphaFold, became the global leader in protein-structure prediction, but it was not sufficiently accurate to be scientifically useful.¹ That system was radically rearchitected by our team, led by John Jumper, and at CASP14 in 2020, AlphaFold 2 essentially solved the grand challenge of predicting protein structure to atomic-level accuracy.² That this breakthrough happened during the COVID-19 pandemic was particularly significant: even in this context, humanity leaped forward and in the process prepared itself with more powerful tools to deal with future pandemics.

Since then, the AlphaFold Protein Structure Database has cataloged predictions for more than two hundred million proteins – nearly every cataloged protein known to science – and made them freely available to researchers everywhere.³ What was once a specialized and often painstaking pursuit was suddenly a navigable landscape, brought into sharp focus by a new scientific lens. AlphaFold 2 revealed what was always present but hidden from view: the structural biology of life, now visible at atomic resolution.

More than three million scientists in 190 countries have already used the AlphaFold database to accelerate their work, saving many millions of dollars and substantial human effort. It is accelerating drug development and advancing scientific progress across a wide range of fields, from combating malaria to discovering enzymes that break down plastic pollution to supporting research to increase honeybees' chances of survival.⁴ AlphaFold is a clear demonstration of how AI can expand access to the tools of discovery, democratizing and accelerating science on a global scale.

Progress did not cease at understanding structures of individual proteins. While AlphaFold 2 essentially decoded the enormous “alphabet” of 3D protein structures, the subsequent breakthrough model, AlphaFold 3, revealed how that alphabet combines into words and sentences in the language of life. The updated 2024 model accurately showed how proteins interact with DNA, RNA, ligands, and each other to carry out coordinated biological processes.⁵ We packaged both models into our free online tool, AlphaFold Server, for the benefit of the global scientific community, following extensive risk assessments and consultation with over fifty experts in biosecurity, research, and industry to ensure safe and responsible deployment.

Yet AlphaFold is not the final word in the understanding of proteins; many more research questions remain open. Scientists on my team at Google DeepMind and around the world are going beyond protein structures to learn to predict how proteins change shape (protein dynamics) and what roles they can play (protein function). AI plays a crucial role in these pursuits and I am confident that we will see many more breakthroughs on these problems.

If mathematics is the language of physics – precise, formal, and universal – then biology speaks in something far messier. It is a language of emergence, interaction, context, and complexity. The Human Genome Project, which allowed us to see our complete genetic code – all three billion base pairs – revealed the complexity of this task.

The genome is our cellular instruction manual. It's the complete set of DNA that guides nearly every part of a living organism, from appearance and function to growth and reproduction. Small variations in a genome's DNA sequence can alter an organism's response to its environment or its susceptibility to disease. But deciphering how the genome's instructions are read at the molecular level – and what happens when a small DNA variation occurs – is among our grandest scientific problems.

AI is the ideal translator for this task, not only because it can go beyond human abilities to learn this language from the massive amount of genetic data that we have collected about living organisms but also because its reasoning abilities can scale to handle the complexity of interactions that result from changes in the genetic code.

My team's own research on this problem led to the development of AlphaGenome and Enformer, new artificial intelligence tools that more comprehensively and accurately predict how single variants or mutations in human DNA sequences impact a wide range of biological processes regulating genes. This was enabled by, among other factors, technical advances allowing the model to process long DNA sequences and output high-resolution predictions. These models can integrate information from across vast stretches of DNA, identifying regulatory relationships previously invisible to researchers and offering a powerful new lens on the mechanisms underlying health and disease.

The genomics team at DeepMind has also leveraged AlphaFold, our protein structure model, to develop models like AlphaMissense, which focuses on the most common type of genetic mutation, whereby a single amino acid in a protein is replaced. The system predicts whether each change is likely to be benign or cause disease. It does this by combining evolutionary conservation – how vital that genetic site has been across millions of years – with structural context, assessing how the change might alter the protein's function. Validated against gold-standard datasets, AlphaMissense outperforms other computational methods, deciphering biology at a scale no human could and translating raw genomic code into functional insight.⁶

AI tools that help us interpret the genome in functionally meaningful ways are increasingly critical. As tools such as the gene-editing technology CRISPR push the boundaries of how we edit and write in the language of life, the ability to merely read DNA is no longer enough. Whether we are designing climate-resilient crops or engineering enzymes to break down plastic waste, the power to edit life depends on the power to understand its underlying code. AI is helping us connect the dots, showing how changes in DNA lead to changes in function, health, and disease.

Researchers are using this same capability to anticipate future health threats. In 2023, at Harvard Medical School, for example, computational biologist Debora Marks and her team developed EVEscape, a generative AI system trained on viral sequences from hundreds of thousands of nonhuman species.⁷ By forecasting which potential mutations are most likely to help viruses evade immunity, EVEscape supports the design of vaccines capable of targeting both circulating strains and those likely to emerge. Just a few years ago, who could have imagined such forward-facing vaccines, designed not just for today's pathogens but for tomorrow's? This marks a fundamental transition in medicine: from a discipline of reaction to one of anticipation.

In 2024, our work on AlphaFold earned Demis and John Jumper, our DeepMind colleague and colead of the AlphaFold project, the Nobel Prize in Chemistry. They shared the award with David Baker, recognized for his contributions to computational protein design. The prize reflects not only the impact of AlphaFold but also a broader shift underway, a wave of AI-driven research transforming discovery across science and industry worldwide.

For example, Baker's team at the Institute for Protein Design has shown that proteins can be designed from scratch as nanoscale machines with programmable shapes and functions: where nature evolves these structures over millions of years, AI-powered tools such as RFdiffusion can now generate and test thousands of candidates *in silico*.⁸ The long-term ambition of Baker's lab is striking: to create bespoke molecular machines that neutralize viruses, repair tissues, catalyze new forms of chemistry, and much more.

That same acceleration is reaching therapeutics. In 2019, Regina Barzilay and her team at MIT used deep learning to discover that a molecule previously researched (and discontinued) as a treatment for diabetes was also a potent new antibiotic active against drug-resistant bacteria.⁹ Identified from a vast digital library of molecules in just hours, the drug, which they renamed halicin, represented the kind of AI-accelerated discovery that is inaccessible or prohibitively difficult using traditional methods.

In neuroscience, AI is visualizing the brain in unprecedented detail. In a landmark collaboration in 2024, Jeff Lichtman at Harvard University and Viren Jain at Google Research used AI to reconstruct a detailed human brain circuit of a small tissue sample. Even though the sample was just half the size of a grain of rice, it required more than a petabyte of electron microscopy data and revealed tens of thousands of cells and over 150 million synapses, including previously unknown structures.¹⁰ Their work heralds a new era in brain science defined by AI-enabled seeing.

So what other breakthroughs should the scientific community pursue? I have never thought of AI as a solution in search of a problem. At Google DeepMind, we begin with a bold question: Which foundational scientific bottlenecks, if solved,

would unlock cascades of transformative, downstream discovery? We call these “root node” problems.

Having defined the target, I ask which of these challenges appear genuinely out of reach – the bottlenecks that scientists broadly agree won’t yield to traditional methods within the next decade. I only want our team to pursue the hardest questions that require the broadest multidisciplinary approaches, the best AI research, and the best compute and data, all coming together to advance the state of the art farther and faster.

Consider materials science and mathematical discovery. Each represents a domain in which progress had slowed, not for lack of effort, but because the sheer complexity of their problems was outpacing available methods. Our strategy is to intervene at these points: assembling multidisciplinary teams, developing bespoke tools, and staying the course until we and the researchers our work supports achieve the necessary breakthroughs. This is not AI as a hammer looking for nails, but as keys carefully designed to unlock the gates of scientific progress.

Materials have shaped our civilizations. From the Stone Age to the Silicon Age, they have powered every major leap in human progress: bronze gave rise to tools and weapons, steel to cities and railways, silicon to the digital world. The next great leap may come from the discovery and deployment of materials with properties we have never seen before. For example, human advancement now depends on us finding new sources of energy and new approaches to efficiently store and transform energy. Many of the most promising approaches – from next-generation solar panels and better rechargeable batteries to room-temperature superconductors – hinge on the discovery of new materials.

Yet materials discovery remains painfully slow. For most of history, it has been driven by trial and error, and the low-hanging fruit was picked long ago. What remains is a largely uncharted search space, too immense for even the most determined researchers to navigate unaided. This bottleneck holds back multiple scientific fields and technologies. But the dam is starting to crack. AI is now changing how materials research is done, shifting away from trial and error and intuition toward prediction and design. AI models can now estimate key properties such as conductivity, stability, and reactivity directly from a material’s atomic structure, having learned from decades of experimental and computational data.

Still, the tens of thousands of scientifically explored stable materials represent only a tiny fraction of the billions of possible atomic arrangements allowed by chemistry. To explore this universe, we need new tools of unprecedented speed and precision. That is why we built an AI system called Graph Networks for Materials Exploration (GNoME).¹¹ Trained on decades of open materials data – including the twenty thousand stable materials already discovered through experimentation and another twenty-eight thousand via conventional computation – GNoME

explores the vast space of potential material structures and accurately predicts the most stable among them.

GNoME was trained to model the laws of quantum chemistry, the fundamental rules that govern how atoms bond and interact. These laws are well understood but very costly to simulate directly. GNoME's breakthrough was in learning to approximate them accurately and at scale, allowing it to quickly scan millions of potential compounds that would have taken centuries to evaluate by traditional means.

It returned with extraordinary results: 2.2 million new materials, 380,000 of which were predicted to be stable enough to potentially power future technologies. Among these were 52,000 new layered compounds akin to the wonder-material graphene and more than five hundred lithium-ion conductors – twenty-five times more than previously reported – that could help improve the performance of rechargeable batteries. We released the stable candidates to the global scientific community, estimating that GNoME delivered the equivalent of eight hundred years' worth of scientific progress that day.

Like a telescope scanning the sky, GNoME mapped a constellation of stable materials, guiding experimentalists toward the brightest regions where breakthroughs are most likely to emerge. In this way, AI has become an invention for invention, a tool for building better tools and making more discoveries, with each spin of the flywheel amplifying the next.

Of course, akin to our early work on protein folding, this is just the beginning. While GNoME can predict potentially stable materials, the most promising of these candidates must also be synthesized and tested in the laboratory and in the real world to verify their properties and potential. While autonomous “self-driving” labs are accelerating this important step, from the A-Lab at Lawrence Berkeley National Laboratory to the Argonne National Laboratory in Lemont, Illinois, we still have further to go before we can fully create bespoke substances *in silico*, tailored to the needs of future technologies.¹²

In materials science, what we need may not yet exist, but the elements to create it are all around us. With AI as our guide, we will begin to conjure new matter from the vast possibility space of nature itself. Yet materials are just one part of building a sustainable future. We also need to pursue every viable path for generating clean, abundant energy. Among the most promising paths – and most difficult – is nuclear fusion, the reaction that powers the stars.

Controlling the superheated plasma inside a fusion reactor is one of science's toughest challenges, requiring precise, rapid adjustments to shifting magnetic fields. In collaboration with EPFL (École Polytechnique Fédérale de Lausanne, home of the Swiss Plasma Center), we developed an AI system that learns to control and shape plasma in real time, opening new paths for fusion research and bringing this clean, abundant energy source closer to reality.¹³

Elsewhere in high-energy physics, detectors at the Large Hadron Collider generate petabytes of data from trillions of collisions, a scale that overwhelms traditional analysis. Physicist Kyle Cranmer at the University of Wisconsin–Madison is doing pioneering work using machine learning to extract rare signals from noisy backgrounds and accelerate the search for new physics.¹⁴ His research exemplifies a growing trend across the physical sciences: combining deep domain knowledge with AI systems that are not just data driven but increasingly aware of physical principles.

Mathematics is the invisible foundation of modern science and technology, from computing and communications to physics, engineering, and medicine. It also is the perfect test bed to study and evaluate the intuition and reasoning abilities of AI systems.

Early large language model (LLM)–based AI systems struggled with even basic mathematical reasoning. Their proclivity to hallucinate made it challenging for them to reason over the long sequences of reasoning steps typically required for nontrivial mathematical problems. Disappointed by these earlier investigations, many experts felt that the mathematical intuition and reasoning required to solve the kinds of complex problems featured at math competitions were many years away.

To overcome these challenges, we built AlphaProof, a specialized agent for solving math problems. Unlike normal LLM-based conversational agents, AlphaProof trains itself to prove mathematical statements in the formal open-source programming language Lean. It couples a pretrained language model with our AlphaZero reinforcement learning algorithm, which previously taught itself how to master the games of chess, shogi, and Go.

Formal languages offer the critical advantage that proofs involving mathematical reasoning can be formally verified for correctness. Their use in machine learning has, however, previously been constrained by the very limited amount of human-written data available. In contrast, natural language–based approaches can hallucinate plausible but incorrect intermediate reasoning steps and solutions, despite having access to orders of magnitude more data. We established a bridge between these two complementary spheres by fine-tuning a Gemini model to automatically translate natural language problem statements into formal statements, creating a large library of formal problems of varying difficulty.

When presented with a problem, AlphaProof generates solution candidates and then proves or disproves them by searching through possible proof steps in Lean. Each proof found and verified is used to reinforce AlphaProof’s language model, enhancing its ability to solve subsequent and more-challenging problems. We trained AlphaProof to compete in the International Mathematical Olympiad, a prestigious competition for young mathematicians and a useful benchmark, by proving or disproving millions of problems, covering a wide range of difficulties

and mathematical topic areas over a period of weeks leading up to the competition. AlphaProof also applied this training loop during the contest, reinforcing proofs of self-generated variations of the contest problems until it could find a full solution.

In 2024, AlphaProof (along with AlphaGeometry, our specialized system for geometry problems) made history by being the first AI system to solve four out of six problems from that year's International Mathematical Olympiad, achieving the same level as a silver medalist in the competition. In 2025, an advanced version of Google's Gemini Deep Think model (which incorporates some findings from AlphaProof) solved five of the problems perfectly, achieving a gold-medal level of performance.¹⁵ This result was repeated by a specialized version of OpenAI's ChatGPT system.

Impressive as it is, solving Olympiad-level problems doesn't alone advance mathematics. To achieve that, AI must produce results that we have not yet discovered. A good example is my team's collaboration with mathematicians to use AI in making significant progress on solving partial differential equations – a key step on the path to solving the Navier Stokes equations – a long-standing fundamental question in fluid dynamics and one of the Millennium Prize Problems.¹⁶ While we are yet to see one of these grand challenge problems in mathematics fall to AI, it is only a matter of time.

A hint at this future for mathematics research comes in the form of our AlphaEvolve system, which can generate novel algorithms and solutions across both computing and mathematics. AlphaEvolve combines the creativity of large language models with automated evaluators to evolve algorithms for math and practical applications in computing. It operates as an evolutionary process at digital speed.

Already, AlphaEvolve has delivered practical gains – improving code and resource efficiency across Google's global computing infrastructure – while also demonstrating capabilities in fundamental mathematics. Tested on more than fifty open problems in areas including geometry, number theory, and combinatorics, AlphaEvolve rediscovered the best-known solutions in 75 percent of cases; in 20 percent of cases, it went further, improving on existing results and advancing the frontier of mathematical knowledge.¹⁷

Crucially, AlphaEvolve does not operate as a black box. It expresses solutions as human-readable programs, providing a transparent chain of reasoning that allows scientists to inspect and build upon its discoveries. In a subsequent study, our team worked together with mathematicians Terence Tao and Javier Gómez-Serrano to evaluate AlphaEvolve on a number of other math problems. They found that AlphaEvolve can help discover new results across a range of problems, including discovering a new construction for the finite-field Kakeya conjecture.

While both AlphaEvolve and AlphaProof have limitations in terms of their efficiency and ability to handle even more complex problems, they signal some of the ways computer scientists and mathematicians will leverage AI to accelerate research.

While AI is already impacting mathematics and science through breakthrough models like AlphaEvolve and AlphaFold, there is a broader transformation now underway: the rise of agentic AI. This new generation of AI systems moves beyond narrow task execution to take a more autonomous role in discovery, not just assisting human inquiry but actively advancing it. Unlike specialist models built for specific domains like protein prediction or weather forecasting, agentic systems are more general and capable of open-ended exploration.

Google's AI co-scientist, built on Gemini 2.0 and launched in 2025, is a case in point. It is a multi-agent system, guided by human-defined goals, that draws on existing literature to generate novel research hypotheses and experimental protocols.¹⁸ I think of it as a symphony of synthetic scientists working to increase the cadence of discovery.

With collaborators including the Fleming Initiative, Imperial College London, and Stanford University, we evaluated AI co-scientist's ability to generate novel hypotheses in biomedicine – including drug repurposing, treatment targets, and antimicrobial resistance – and validated these hypotheses through laboratory experiments. The results were promising, suggesting that AI will help shape the research agenda itself by surfacing salient questions before anyone thinks to ask them. Future systems will be increasingly general. They will propose their own hypotheses, design and refine experiments, and seek out the data they need, perhaps even navigating the physical world to gather it (with human oversight, of course).

We are moving toward AI that doesn't just facilitate science but begins to *do* science. For now, however, our human collaborators face a specification problem. They are finding that as these hypotheses-generating AI agents become more powerful, it is increasingly important for human scientists to be extremely precise in specifying the problem they want these agents to solve. Otherwise, the AI agents tend to find a solution that violates assumptions the human scientists were making but had not included in the original problem description. The burden of reasoning is shifting; we are moving from being the solvers of puzzles to the architects of questions.

This last hurdle foreshadows a future question: What happens when our tools begin to uncover truths we struggle to grasp? Science has always aimed not only to predict but to understand – to build coherent theories that reveal how the world works. Yet as AI grows more capable, it may pose questions or offer insights that resist easy explanation or even human comprehension. So our task as scientists will evolve from interpreting what AI finds to understanding why it matters and acting on it responsibly.

Until now, science has been a strictly human endeavor. The discovery of the scientific method and its focus on dispassionate hypotheses set humanity on the course to modernity. We now stand at an inflection point; the tools we built using the scientific method are promising to propel us into a vast new unknown for the first time in centuries.

As humanity moves ahead and draws the first maps of this new world, we must tread carefully. Long term, if we become dependent on systems we do not fully understand, we risk accepting their answers without sufficient scrutiny. More than a century ago, E. M. Forster's *The Machine Stops* imagined a future in which humanity lives passively under a vast machine's care, having long forgotten how the world – and the Machine – works.¹⁹ As we build future generations of AI scientific assistants, we must remain clear-eyed about what we are gaining – and what we could lose along the way.

Given the pace of progress, making predictions more than even a few months ahead feels like a fool's errand; but I can say with confidence that the breadth and scale of advances AI will soon unlock are difficult to overstate. Within five years, AI will have enabled dozens, if not more, Nobel Prize-level breakthroughs. This is what happens when you give scientists superpowers.

AI will tackle problems that have resisted human insight for decades, from the deepest questions in mathematics (such as the Millennium Prize Problems) to the design of high-temperature superconductors that could radically improve how we generate and transfer energy. In biology, future AI systems may simulate entire cells, design synthetic organisms with myriad applications, and eliminate entire categories of disease.

Of course, the pace of progress is really the pace of understanding. For most of human history, new understandings took hold across centuries, from Copernicus to Newton, Darwin to Watson and Crick. Over time, that cadence has quickened, from decades to years to – now, with AI – months or even days. As our new scientific tools collapse the distance between question and answer, we find ourselves at an inflection point not just in capability but in tempo. And for those of us building these tools, we may yet have farther to go, but it feels like we are on the most exciting scientific journey ever undertaken.

At the precipice of change, some have wondered, “When does a machine win the Nobel Prize? And after that, will humans ever win it again?” These questions capture the common anxiety that AI might one day eclipse human contributions altogether. This is understandable but misses the point of why we do science. We do not just seek answers; we seek meaning. In that sense, AI will not replace scientists. Instead, it will expand what they can imagine, understand, and achieve. After all, the telescope didn't make astronomers obsolete. It gave them the stars.

ABOUT THE AUTHOR

Pushmeet Kohli is the Chief Scientist of Google Cloud and Vice President of Science at Google DeepMind. He previously served as Director of Research at Microsoft Research. He has recently published in *Nature*, *Science*, *Nature Medicine*, and *Advances in Neural Information Processing Systems*.

ENDNOTES

- ¹ Andrew W. Senior, Richard Evans, John Jumper, et al., “Improved Protein Structure Prediction Using Potentials from Deep Learning,” *Nature* 577 (7792) (2020): 706–710.
- ² John Jumper, Richard Evans, Alexander Pritzel, et al., “Highly Accurate Protein Structure Prediction with AlphaFold,” *Nature* 596 (7873) (2021): 583–589.
- ³ Ewen Callaway, “‘The Entire Protein Universe’: AI Predicts Shape of Nearly Every Known Protein,” *Nature*, July 29, 2022, <https://www.nature.com/articles/d41586-022-02083-2>.
- ⁴ “Stopping Malaria in Its Tracks,” Google DeepMind, October 13, 2022, <https://deepmind.google/discover/blog/stopping-malaria-in-its-tracks>; “Creating Plastic-Eating Enzymes That Could Save Us from Pollution,” Google DeepMind, July 28, 2022, <https://deepmind.google/discover/blog/creating-plastic-eating-enzymes-that-could-save-us-from-pollution>; and “How the Honeybee Could Help Protect Species around the World,” Google DeepMind, July 28, 2022, <https://deepmind.google/discover/blog/how-the-honeybee-could-help-protect-species-around-the-world>.
- ⁵ Josh Abramson, Jonas Adler, Jack Dunger, et al., “Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3,” *Nature* 630 (8016): 493–500.
- ⁶ Jun Cheng, Guido Novati, Joshua Pan, et al., “Accurate Proteome-Wide Missense Variant Effect Prediction with AlphaMissense,” *Science* 381 (6664) (2023), <https://doi.org/10.1126/science.adg7492>.
- ⁷ Nicole N. Thadani, Sarah Gurev, Pascal Notin, et al., “Learning from Prepandemic Data to Forecast Viral Escape,” *Nature* 622 (7984): 818–825.
- ⁸ Joseph L. Watson, David Juergens, Nathaniel R. Bennett, et al., “De Novo Design of Protein Structure and Function with RFdiffusion,” *Nature* 620 (7976) (2023): 1089–1100.
- ⁹ Jonathan M. Stokes, Kevin Yang, Kyle Swanson, et al., “A Deep Learning Approach to Antibiotic Discovery,” *Cell* 180 (4) (2020): 688–702.
- ¹⁰ Alexander Shapson-Coe, Michał Januszewski, Daniel R. Berger, et al., “A Petavoxel Fragment of Human Cerebral Cortex Reconstructed at Nanoscale Resolution,” *Science* 384 (6696) (2024), <https://www.science.org/doi/10.1126/science.adk4858>.
- ¹¹ Amil Merchant, Simon Batzner, Samuel S. Schoenholz, et al., “Scaling Deep Learning for Materials Discovery,” *Nature* 624 (7990) (2023): 80–85.
- ¹² Michael Eisenstein, “Self-Driving Laboratories, Advanced Immunotherapies and Five More Technologies to Watch in 2025,” *Nature* 637 (8047) (2025): 1008–1011; Lauren Biron, “Meet the Autonomous Lab of the Future,” News from Berkeley Lab, April 17, 2023, <https://newscenter.lbl.gov/2023/04/17/meet-the-autonomous-lab-of-the-future>; and “Autonomous Discovery: Reimagining the Future of Scientific Discovery Through

- AI, Robotics and Labwide Collaboration,” Argonne National Laboratory, <https://www.anl.gov/autonomous-discovery> (accessed December 18, 2025).
- ¹³ Jonas Degraeve, Federico Felici, Jonas Buchli, et al., “Magnetic Control of Tokamak Plasmas through Deep Reinforcement Learning,” *Nature* 602 (7897): 414–419.
- ¹⁴ The ATLAS Collaboration, “An Implementation of Neural Simulation-Based Inference for Parameter Estimation in ATLAS,” *Reports on Progress in Physics* 88 (6) (2025): 067801.
- ¹⁵ Thang Luong and Edward Lockhart, “Advanced Version of Gemini with Deep Think Officially Achieves Gold-Medal Standard at the International Mathematical Olympiad,” Google DeepMind, July 21, 2025, <https://deepmind.google/discover/blog/advanced-version-of-gemini-with-deep-think-officially-achieves-gold-medal-standard-at-the-international-mathematical-olympiad>.
- ¹⁶ Yongji Wang and Sam Blackwell, “Discovering New Solutions to Century-Old Problems in Fluid Dynamics,” Google DeepMind, September 18, 2025, <https://deepmind.google/blog/discovering-new-solutions-to-century-old-problems-in-fluid-dynamics>; and Minglang Yin, Nicolas Charon, Ryan Brody, et al., “A Scalable Framework for Learning the Geometry-Dependent Solution Operators of Partial Differential Equations,” *Nature Computational Science* 4 (12) (2024): 928–940.
- ¹⁷ Alexander Novikov, Ngan Vũ, Marvin Eisenberger, et al., “AlphaEvolve: A Coding Agent for Scientific and Algorithmic Discovery,” arXiv (2025), <https://doi.org/10.48550/arXiv.2506.13131>.
- ¹⁸ Juraj Gottweis and Vivek Natarajan, “Accelerating Scientific Breakthroughs with an AI Co-Scientist,” Google Research, February 19, 2025, <https://research.google/blog/accelerating-scientific-breakthroughs-with-an-ai-co-scientist>.
- ¹⁹ E. M. Forster, “The Machine Stops,” *Oxford and Cambridge Review* 8 (1909).

Beyond Representation: AI in Cellular Discovery

Charlotte Bunne & Aviv Regev

A new paradigm is emerging at the intersection of artificial intelligence and experimental biology, where cells are no longer merely observed, but comprehensively modeled, queried, and predicted in silico. New measurement technologies and the ability to genetically manipulate cells precisely have opened the way to measure cells in extraordinary detail under tens of thousands of perturbations. Concomitantly, AI foundation models are learning to represent, simulate, and even anticipate cellular behavior. This essay traces the convergence of these revolutions, showing how they are giving rise to “virtual cells”: integrative models that unify diverse molecular and spatial data into coherent, functional representations that can generalize across biological contexts and conditions. Beyond representing and interpreting biological lab measurements, virtual cells aim to predict unseen outcomes, imagine new contexts, and guide discovery. In closing the loop between data generation and hypothesis testing, AI is transforming biology into a self-refining, interactive science, resulting in a profound shift: from observing life to actively modeling it, with implications for precision medicine, biotechnology, and the scientific method.

Understanding the immense complexity of biology requires characterizing life at its fundamental units – cells – in all their diversity of molecules, behaviors, and contexts. Humans are composed of an estimated thirty-seven trillion cells spanning many thousands of specialized types and dynamic states, each influenced by its genetics, current environment, and lived history. While cells within an individual all carry approximately the same set of instructions in their DNA, they operate by using (“expressing”) only specific components at a time (transcribing into RNA and translating the RNA into protein), as needed for their specific role and temporal context in the body, and thus vary substantially in their molecular characteristics and the functions they can execute across types, times, and conditions.

Deciphering this cellular kaleidoscope is essential for both basic science and medicine. Every process of human life, from brain activity to immune defense, emerges from the collective actions of diverse cells. Errors in these processes underlie disease: autoimmune conditions reflect immune cells turning against self-tissues, cancers emerge from a single cell with a mutated genome that ignores regula-

tory signals, and neurodegeneration occurs when vulnerable neuronal subtypes die because of the accumulation of toxic molecular aggregates and inappropriate immune cell activity. Understanding the molecular programs and interactions of cells is therefore the foundation for diagnosing disease, identifying therapeutic targets, and designing effective interventions.

Despite this cellular variation and its importance, molecular studies of cells have historically averaged across many of them at once, largely due to the limitations of measurement technologies. RNA levels, for example, were typically measured from cell cultures or tissues (and thus from many cells and not individual *single* cells) ground up into a molecular “smoothie.” While powerful, these bulk approaches hid the rich cell-to-cell heterogeneity that defines cells in both health and disease. Thus, a tumor sample’s average gene readout might miss a rare drug-resistant cell subpopulation; likewise, an immune organ’s bulk profile might overlook a few critical regulatory cells amid thousands of others.

Groundbreaking technological advances over the past decade – combining new lab measurements with powerful AI algorithms – have led to a historic shift from this aggregate view to a single-cell resolution of life. Innovations in single-cell sequencing and cytometry now allow researchers to profile millions of individual cells for their genomes, transcripts, proteins, and more – one cell at a time.¹ Early single-cell RNA sequencing (scRNA-seq) experiments revealed astonishing diversity even in what were thought to be uniform cell populations, overturning more categorical classifications of cell “types” and uncovering new cell states (for instance, previously unrecognized subtypes of neurons or complex, continuous immune cell states).² This laid the foundation for the Human Cell Atlas (HCA) initiative launched in 2016, which set out to create comprehensive reference maps of all human cells as a basis for both understanding human health and diagnosing, monitoring, and treating disease.³ Like a “Google Maps of the human body” that chart diverse landscapes at multiple levels of resolution and across different modalities (topography, street map, traffic patterns, and more), the Human Cell Atlas maps the multitude of cell types and states across tissues, organs, and systems in terms of different molecular and visual properties. By 2017, the HCA and similar efforts had demonstrated that it was possible to discover and catalog myriad cell types across the body using large-scale single-cell profiling; as of 2025, “version 1” atlases of eighteen organs and systems are near-complete.⁴ This massive data-gathering phase proved that a “census of cells” in the human body was within reach. In parallel, experimental strategies such as Perturb-seq began to extend the atlas concept beyond description: Perturb-seq couples pooled CRISPR-based gene perturbations with single-cell RNA-seq so that the effect of turning genes off or on can be read out for thousands of cells at once.⁵ In this way, Perturb-seq systematically links genetic changes to complex cellular responses, producing causal datasets that complement the observational maps of the atlas.⁶

Moving to single-cell genomic resolution fundamentally changed how we think about and interrogate biological systems and led to a plethora of basic scientific discoveries and medically relevant advances.⁷ We now focus on tissues as ecosystems of distinct cells interacting and changing over time within their native contexts. In cancer, for example, single-cell studies revealed a tumor and its microenvironment as an evolving mosaic of malignant and infiltrating immune cells, each with its own gene expression patterns. In neuroscience, single-neuron analyses have exposed hundreds of new subclasses of brain cells that were invisible in bulk measurements.⁸ In human development, the continuous processes by which the fertilized egg leads to thousands of different cell types could be recovered, showing far less discrete determinism than previously assumed. Across many organs, cell types as rare as 1 percent or 0.1 percent of the tissue cells have emerged as crucial nodes in their functioning, from the rare ionocytes in the lung and airways to enteric neurons in the gut.⁹ These insights also have direct medical implications: Single-cell atlases that link disease-associated genes to the specific cell types and contexts in which they act are clarifying mechanisms and guiding intervention. This enables more precise diagnostics, the discovery of novel drug targets, and the development of therapies across cancer and infectious, rare, and degenerative diseases.¹⁰

However, early single-cell methods typically captured only one layer of information at a time (most typically, RNA transcripts), dissociated cells from their native tissue context, and focused on the observed cells rather than the biological mechanisms that give rise to and control them. To fully understand cellular function, scientists have had to push further, combining multiple modalities per cell, deciphering the underlying causes of their actions, and putting them back into the spatial context from which they come.

Biology is inherently multidimensional, multimodal, and dynamic. A cell's identity is not defined by only one feature (say, a specific gene) or one category or modality of features (say, its mRNA levels) but by a constellation of many features (multidimensional) or many categories (multimodal) – DNA sequence and epigenetic marks, RNA transcripts, proteins, metabolites, physical shape at multiple length scales, location in a tissue, and interactions with neighbors – all of which change in time and space.¹¹

Recognizing this, the field has rapidly progressed through advances in technological measurements and computational algorithms from “snapshot” single-modality single-cell measurements to multimodal assays in cells, including dynamic ones. Today, it is feasible to profile multiple molecular layers from the same cell, for example, simultaneously capturing a cell's transcriptome (which genes are active) and its epigenome (how the DNA is marked and organized), or measuring both RNA and proteins.¹² Both experimental approaches and computational algorithms help recover dynamics, either by tracking live cells directly or by inferring dynamics from

static snapshots or molecular recorders and lineage tracers.¹³ These multimodal methods provide a more holistic view of each cell, linking gene regulation to gene expression to protein output. Instead of a one-dimensional profile, we get a richer “fingerprint” of each cell’s state, its location, and even its history and potential future.

In parallel, a revolution in spatial biology is recovering the two- and three-dimensional (2D/3D) context that traditional single-cell profiling lost. New spatial transcriptomics technologies allow us to map molecular information in 2D tissue sections and even 3D whole organs.¹⁴ Lab methods can now pinpoint individual RNA or protein molecules in cells at micron-scale resolution, whereas other techniques can recover the 2D location of a cell prior to profiling.¹⁵ The result is a tissue map showing what cell types are present and what they express, how they relate to histological images, and what tissue structures and other cells they touch. This information is crucial because cells’ behaviors and identity are profoundly influenced by their neighbors and proximal (micro) environment: a T cell nestled next to and precisely recognizing a tumor cell as “non-self” may act differently than the same T cell elsewhere; a neuron’s role in a circuit depends on its location and connections. Notably, profiling methods now span multiple scales, from subcellular imaging of organelles to cellular-resolution tissue maps to whole-organ scans. Thus, the rise of spatial profiling and advanced imaging enables us to fulfill the vision of the Human Cell Atlas not just through a cell census, but in a 3D, multiscale map of cells in the body.¹⁶

Moving from one-dimensional catalogs to multimodal, spatially resolved maps also shifts us from a cell biology centered on how cells are organized and function to a tissue biology focused on the holistic and integrated organization and function of tissues. We can begin to ask questions that were previously far less accessible: How do cells arrange into niches and communities? Which cell-to-cell interactions drive responses to injury or infection? How do signals propagate through the spatial architecture of a tissue, and how are they coordinated to perform the tissue’s and organ’s ultimate function? However, to address each of these questions requires generating, analyzing, and reasoning over enormous volumes of complex data. Indeed, it demands more than traditional analysis by calling for a partner that can navigate immense dimensionality, link disparate measurements, and surface hidden patterns and their underlying causes at scale.

This partner is artificial intelligence, which now serves as an essential component for making sense of the vast, multidimensional datasets produced by modern single-cell and spatial profiling. As experimental methods have scaled, from bulk assays to high-throughput and multi-omic measurements, machine-learning algorithms have matured. Indeed, this was not merely a fortunate coincidence but a mutually reinforcing dynamic: large volumes of rich, heterogeneous data fuel advances in AI architectures, while more capable AI tools in turn inspire and enable ever more ambitious experimental designs. Techniques such as representation learning com-

press tens of thousands of cellular features into biologically meaningful embeddings, enabling the discovery of hidden subpopulations and trajectories. Droplet-based single-cell RNA-seq, for instance, captures only 5 to 20 percent of each cell's transcripts, producing sparse and noisy profiles.¹⁷ Here, representation learning was essential to recover structure and make the method transformative.¹⁸ In parallel, integration frameworks aim to align disparate modalities – transcriptomics, proteomics, epigenomics, and imaging. These algorithms reconcile technical noise, resolve batch effects, and synthesize partial measurements so that each cell's composite profile emerges more coherently than any single assay could provide. They can also perform cross-modal inference: for example, predicting gene-expression patterns directly from histological images, thereby extending molecular insight to routine clinical specimens.¹⁹ Beyond these applications, AI contributes through predictive and generative modeling, allowing annotation of complex datasets and forecasting of cellular behavior under unseen conditions. Such models can anticipate how cells respond to perturbations or therapies or reconstruct dynamic processes from destructive measurements.²⁰

One of the most significant developments of the convergence of AI and biology is the emergence of foundation models for cell biology. In AI, “foundation models” are very large models trained on massive unlabeled datasets to perform a wide range of tasks that extend well beyond those used for pretraining.²¹ At the core of these foundation models lies the principle of self-supervised learning, a training approach that leverages inherent patterns in the data without relying on explicit labels provided by researchers. The first successful demonstration of foundation models was in natural language processing, where self-supervised methods train models by predicting masked or omitted components of text or by generating the next syllable in a sentence.²² The performance of the resulting large language models (LLMs) has been astonishing, transforming countless professional workflows – from writing, summarizing, and coding to research and customer service – and reshaping how individuals engage with and consume information in everyday life. Extending this paradigm to cell biology follows naturally, as large-scale experimental biology provides precisely the kind of data environment in which self-supervised AI thrives. High-throughput single-cell and spatial experiments routinely generate datasets of millions of cell profiles, each with thousands of molecular and spatial features, creating exactly the rich and diverse input that cell atlas foundation models need to learn robust, generalizable biological representations.²³ With this approach, scientists aim to mirror the success of LLMs, which derive powerful predictive capabilities by ingesting vast quantities of text data. Thus, where early single-cell genomics representation learning compressed noisy measurements into descriptive embeddings, foundation models are now trained on diverse, heterogeneous inputs across scales and modalities to

capture the underlying grammar of cellular systems: representations that should support prediction, querying, and generation well beyond the measured data.²⁴

However, biological datasets present distinct challenges. They are inherently multimodal (integrating transcriptomics, proteomics, epigenomics, and visual measurements) and multiscale (bridging molecular details to tissue-level organization and spanning nano, micro, and macro scales). Moreover, they represent dynamic biological systems captured through static snapshots, and they are governed by genetic and molecular causes. Addressing this complexity requires new computational architectures and training strategies specifically tailored to biological data types.

First, operationalizing large-scale foundation models within a cellular context requires reformatting raw experimental data into representations that algorithms can readily process. Experimental measurements are decomposed into standardized units, analogous to the “tokens” (basic units of text) in language models, that capture molecular features, spatial coordinates, and other cellular attributes. These tokenized profiles, derived from transcriptomic, proteomic, epigenomic, and spatial assays, are then processed by transformer-based architectures.²⁵ Such models learn statistical relationships – or a “grammar” – among these biological tokens across large cohorts of cells.

Next, the way in which these models are trained determines what kinds of patterns they capture. One family of self-supervised approaches uses masking, whereby a subset of features within a profile is hidden and the model must reconstruct them from the remaining context. This forces the model to capture covariation among genes and molecules and has proven highly effective in large single-cell and tissue foundation models.²⁶ A second strategy is integration-based alignment, whereby paired measurement types are co-embedded so that consistent structure across modalities reinforces a unified cellular representation. This allows, for instance, protein-localization patterns observed in imaging to be aligned with protein-interaction measurements.²⁷

A third novel strategy is cross-modal prediction, wherein the model is given one modality and asked to generate another. Tasks such as predicting protein abundance from RNA levels or inferring gene expression from histology images force the model to learn systematic correspondences across measurement spaces and enable translation between them.²⁸ In two other approaches, contrastive learning methods bring together representations of identical states and separate those of distinct states, structuring the embedding space so that distance reflects biological similarity rather than technical variation, while generative objectives require them to produce realistic new profiles that obey biological constraints.²⁹ Together, these objectives provide strong training signals, though it remains an open question whether they fully capture the biological rules at play, leaving room for the discovery of new learning objectives to train multiscale and multimodal foundation models for cellular biology.

Like other foundation models, once trained, these models can be applied to various downstream tasks distinct from those used for training. Using simple (linear) classification or regression models trained on the learned representations, they can distinguish subtle cellular states or predict quantitative responses, such as estimating how strongly a cell will activate under a given stimulus.³⁰ In retrieval tasks, their embeddings enable efficient matching, such as linking a rare or poorly characterized cell in a new dataset to its closest counterparts across large reference atlases, or identifying similar patients whose profiles can inform clinical decision-making.³¹

Beyond tokenization and learning objectives, the development of foundation models for cell biology has driven the design of novel architectures that move beyond the standard transformers used in language and vision. Whereas language models operate on sequential tokens and vision models on grid-structured pixels, biological data inhabit irregular, high-dimensional spaces: genes interact through complex regulatory networks, cells reside in continuous spatial and temporal contexts, and measurements are frequently sparse or missing. Meeting these demands has, for example, led to specialized attention modules, such as operators that extend context windows to capture long-range genomic interactions across millions of bases, and mechanisms that scale to the extreme dimensionality of spatial omics. In parallel, hierarchical transformer designs represent tissues across multiple levels of organization, linking molecules within cells, cells within microenvironments, and microenvironments within tissues. Another direction develops architectures that operate directly on sets of cells or continuous spatial and temporal coordinates, enabling models to capture local neighborhoods and smooth dynamics without imposing artificial sequence or grid structures. Strategies have also been introduced to ensure robustness across heterogeneous assays, such as technologies that capture thousands of RNA molecular species but only a limited number of protein species.

Advances in the design of foundation models for cell biology have already yielded practical scientific benefits. Models trained in a self-supervised manner on hundreds of millions of single-cell profiles – a scale of data integration across cell types, tissues, and organs not previously attempted in computational biology – outperform classical tools in recognizing cell types, identifying gene functions, and reconstructing gene interaction networks.³² Their embeddings also make it easier to pinpoint rare subtypes.³³ Another striking capability comes from models that integrate microscopy images with molecular profiles. Traditionally, pathologists could only qualitatively infer cellular activity from a tissue slide, but AI is changing that. Recently developed foundation models link histology to gene expression: by training on over 2.2 million paired images and RNA measurements from the same tissue “pixel,” the model learned to predict which genes are active in a section directly from an H&E (Hematoxylin and Eosin)-stained image.³⁴ In practice, this means we can take an ordinary histopathology slide of a patient’s biopsy and approximate the spatial pattern of gene activity – without any specialized measurement.

Foundation models are also beginning to address causality: For example, by identifying genes whose action controls the cellular phenotypes, like RNA levels. Lab experiments typically rely on genetic manipulations, changing genes one at a time (knockout or overexpression) and observing the effect. Single-cell genomics methods combined with CRISPR-based perturbations, as in Perturb-seq, allow thousands of genetic changes to be tested simultaneously, each cell carrying a different edit and being read out at single-cell resolution.³⁵ Even so, the combinatorial space of possible perturbations remains far larger than what can be experimentally tested. Foundation models offer a path forward by predicting perturbation outcomes, a key step toward true virtual experiments. In recent studies, models trained on large compendia of unperturbed cells were fine-tuned to anticipate how those cells' transcriptomes would shift after specific interventions (such as a drug or a CRISPR-based gene knockout).³⁶ These models generalized to new, unseen perturbations with partial success, effectively performing zero-shot predictions.³⁷ For example, without ever having observed a particular drug in training data, a model could predict with reasonable accuracy which genes increase or decrease because of it, and by how much, in a given cell type, extrapolating from patterns learned during broad pretraining. While still evolving, early results show that these models often match or exceed the accuracy of simpler predictive methods. Moreover, they do so in a mechanism-agnostic way: the model is not explicitly told about molecular pathways or networks; it infers likely cascades of change from data. This ability to successfully model causal effects in a cell or, essentially, to perform *in silico* perturbation experiments would represent a fundamental advance in computational biology.

Beyond foundation models, the convergence of AI and high-throughput cell biology is now entering a transformative new phase in which artificial intelligence moves past passive analysis to actively participate in, and even drive, the scientific discovery process. The next frontier involves developing AI agents: autonomous systems that can actively orchestrate entire research workflows. Unlike traditional AI models that passively analyze data or make predictions based on a user's prompt, an AI agent is a system that can perceive its environment, make decisions, and take actions to achieve specific goals. In the context of biological research, these agents can act as intelligent intermediaries between computational models and laboratory experiments, essentially becoming digital research assistants that can propose hypotheses, design experiments, and learn from results.³⁸

AI agents are beginning to transform the traditional scientific method through a lab-in-the-loop approach – a paradigm in which the agent integrates AI models, multiple analysis tools, and experimental platforms in a closed-loop cycle. Using this approach, it might propose a hypothesis or drug candidate based on computational predictions from a model, instruct robotic lab instruments to execute the experi-

ment, analyze the results, and then use those findings to refine its next hypothesis, all with minimal human intervention. This creates a highly iterative and adaptive workflow that has the potential to dramatically accelerate the pace of discovery. Rather than the traditional linear progression of design → experiment → analysis, the lab-in-the-loop paradigm enables massively parallel experimentation with continuous feedback loops, in which computational predictions and experimental results are seamlessly integrated at unprecedented scales.

Central to achieving this integration are sophisticated AI technologies that enable both virtual experimentation and intelligent decision-making. First, the AI agents need to be powered by effective generative models: machine-learning systems trained to learn the underlying patterns and distributions of complex biological data, enabling them to generate new, realistic biological scenarios that have never been observed.³⁹ Unlike traditional predictive models that merely classify or forecast outcomes, generative models can create entirely novel molecular structures, cellular behaviors, or experimental designs by sampling from learned biological distributions. In the lab-in-the-loop context, these models serve as powerful hypothesis-generating engines, simulating thousands of potential experiments *in silico* before any physical resources are consumed. For instance, an AI agent might use a generative model to explore how modifying specific nucleic acids in a regulatory DNA sequence could enhance or repress the expression of a gene, effectively running virtual experiments across a vast design space to identify the most promising candidates for laboratory validation.⁴⁰ Importantly, such closed-loop experimentation can be pursued with two distinct goals: either to address a concrete biological question, such as identifying a more stable protein variant, or to broaden and refine the generative model itself so that it captures biological principles more universally. Once the agent's suggested experiment is executed in the lab, the outcome data are immediately fed back into the system, enabling the generative model to refine its understanding of the biological landscape and improve future predictions.

Second, to fully realize the potential of autonomous scientific discovery, future lab-in-the-loop systems will need to incorporate reinforcement learning, a machine-learning paradigm in which agents learn optimal strategies through trial and error, receiving rewards or penalties based on the outcomes of their actions.⁴¹ Unlike supervised learning that requires labeled examples, reinforcement learning excels in sequential decision-making scenarios in which the best path forward must be discovered through exploration. In the envisioned framework, such systems would treat each experiment as an “action” and assign feedback based on the experimental outcome. An experiment yielding results that confirm a hypothesis or improve a target metric – such as enhanced enzyme activity or drug efficacy – would generate a positive reward, prompting the agent to update its strategy accordingly. This approach could enable AI agents to systematically navigate the vast experimental landscape (which cannot ever be fully tested in a lab), gradually learning optimal

experimentation policies that maximize scientific insight while minimizing time and resources. Maximizing the potential of reinforcement learning in this setting will require shifts in biological experimentation. Increased automation, miniaturization, and continuous monitoring and measurements can all greatly accelerate the success of this scientific process. Moreover, while large-scale experimental investments in biology were historically conducted in batch mode (one large experiment conducted to completeness prior to analysis), in a reinforcement learning setting, it is preferable to use the same experimental resources gradually across iterations. Finally, to reap the broader benefit for the model, scientists may need to accept that a given cycle may not provide the optimal experiment to address a specific question of interest.

This integration of AI agents with experimentation is particularly transformative when applied to the massive datasets generated by modern single-cell and multi-omics technologies. Although researchers can still design informative studies, systematically navigating the combinatorial space opened by datasets that profile millions (and more) of single cells across several molecular layers is beyond what unaided human reasoning can optimize efficiently. AI agents, on the other hand, thrive on such complexity. For instance, when analyzing a single-cell transcriptomic dataset from a tumor sample, an AI agent could identify rare subpopulations characterized by distinctive molecular signatures based on integrated pathway databases, prior biological literature, and predictive modeling of gene interactions. It might then use generative models to simulate the functional outcomes of perturbing hundreds of candidate regulatory genes *in silico*, selecting the most informative targets predicted to have the greatest impact on the cell phenotype or on the tissue niche. These selected targets can then be tested experimentally through high-throughput, multiplexed perturbation screens executed in cells in culture or in an animal model. Upon completion of these experiments, the AI agent could instantly integrate these new measurements, update its predictive models of the underlying regulatory networks, and refine subsequent experimental designs. By leveraging vast molecular knowledge, high-dimensional inference, and real-time feedback, such AI-driven workflows could help systematically identify critical biological insights far more efficiently than manual methods – either autonomously or collaboratively with a scientist user – maximizing the learnings achieved from a given volume of experiments and resources.

These capabilities are now converging toward an ambitious vision: virtual cells.⁴² Virtual cells are computational representations that integrate diverse molecular, spatial, and temporal data to capture, generalize, and simulate cellular behavior and dynamics. A virtual cell is a digital twin of a biological cell, and can be probed and perturbed in the computer just as a real cell might be in the lab. The prospects of this approach are intriguing. If successful, scientists could

predict a cell's response to virtually any change (a drug, a genetic mutation, a viral infection) without needing to perform every experiment in a dish. Much as LLMs learned to “decode” and “speak” natural language, these biological models aim to decode the experimental language of cells, enabling us to virtually “ask” a cellular system what would happen if we tweak a gene or expose it to a new stimulus.

The foundation models describe first steps toward but not yet realizations of the virtual cell approach. Their ability to capture cellular identities, predict perturbation outcomes, and bridge imaging with molecular data forms the scaffold on which more comprehensive virtual cells will be built.⁴³ At the same time, AI agents are emerging as an operational layer that can interrogate these models, generate hypotheses, design targeted experiments, and orchestrate model improvements, creating an interactive loop between *in silico* insight and *in vitro* validation.⁴⁴ Together, agents and foundation models hint at a future in which we can maintain computational mirrors of cells that behave with realistic biological fidelity and are continuously refined by experimental feedback. The concept of virtual cells thus encapsulates this momentum: with sufficient data, sophisticated learning, and agent-driven orchestration, we can create representations that capture a cell's essential behaviors and responses. Although such pioneering efforts are still in their early stages, they offer a glimpse of what is coming. If realized, virtual cells would make it possible to explore biological questions at a scale that no laboratory can match, rapidly screening interventions, uncovering mechanisms hidden from direct measurement, and tailoring therapies to patient-specific cellular states.

We stand at a moment akin to the advent of the microscope, but now empowered by silicon, molecules, and statistics. Just as seventeenth-century optics revealed an invisible world of cells and microbes, today's single-cell and spatial multi-omics analysis combined with AI is exposing hidden patterns, interactions, and potentials within and between cells. The relationship is synergistic and catalytic; high-throughput methods generate rich, multidimensional datasets, and AI transforms those data into coherent models, uncovering hidden patterns and pointing to the next questions to explore. Absent machine learning and artificial intelligence, single-cell genomics methods would have yielded uninterpretable data. Absent single-cell genomics, the data for learning AI models of molecular cell biology would not have been available at the needed scale. The notion of a virtual cell encapsulates this combination: a model trained on large collections of multiscale and multimodal measurements, so faithful and comprehensive that it can serve as an *in silico* surrogate for the real biological system.⁴⁵

The consequences of this shift will extend across science and medicine. In the clinic, diagnostics will move from organ-level descriptions toward pinpointing specific cellular subpopulations and the molecular interactions underlying disease. With models that translate and generate across modalities, even a simple blood

draw might be transformed into a comprehensive multi-omic profile, revealing spatial and functional states of cells that previously required invasive biopsies.⁴⁶ Therapeutic strategies, from small molecules to engineered cells, will increasingly rely on predictive models that simulate cellular responses and generate mechanistic explanations in partnership with AI agents.⁴⁷ Beyond medicine, similar approaches promise to illuminate how cellular ecosystems adapt in evolution and ecology. At a deeper level, these developments may alter our very conception of biology, shifting us from reductionist perspectives to viewing life as a complex, dynamic information system that we can systematically decode, rationally manipulate, and predict.

Many challenges remain, and humility is essential. Biology abounds with context-dependent phenomena, nonlinear feedback, multiple physical and temporal scales, many variables that are challenging to measure, and stochastic noise that confound even the most sophisticated models. Foundation models trained on current datasets may absorb technical artifacts or sampling biases, leading to predictions that fail to generalize across tissues, laboratories, or patient populations. Many molecules, structures, and interactions remain exceedingly difficult to measure, especially in three dimensions and over time; and in most demanding settings, models still cannot match laboratory experiments. Moreover, the “black box” character of deep architectures raises questions about interpretability and trust: without clear mechanistic explanations, it is hard to distinguish genuine insight from spurious correlation. To build confidence, models must undergo rigorous validation, reproducing results across cohorts, predicting outcomes of prospective perturbations, and withstanding real-world application.

Further, biological machine learning has so far focused primarily on prediction: identifying cell types, forecasting perturbation outcomes, or inferring spatial arrangements. What is needed next is the ability to query models in ways that surface concepts, much as one might prompt a colleague by asking not only what will happen but why. Beyond queryability lies reasoning, where models can weigh competing causal explanations, such as whether an immune response is driven by local cytokine gradients or by signals from distant tissues and what may be the functional reasons or consequences. Finally, true transformation will require long-term thinking: systems that can carry a biological problem over weeks or months, continuously refining hypotheses in the background, like a scientist who keeps returning to a puzzle until a unifying explanation emerges. Together, these capabilities would turn artificial intelligence from a predictive tool into a co-investigator, capable of uncovering mechanisms and reshaping how biology is practiced.

The path forward depends as much on advances in experimental biology as on algorithms. New biotechnologies are required to make the invisible measurable, such as molecules, states, and interactions beyond current reach, and to capture multiple modalities simultaneously within the same cell. Prediction alone is insufficient. True understanding demands deep knowledge of the biological system,

careful analysis, and often small-scale validation experiments that cannot necessarily be scaled away. Experimental biology therefore remains central. In this space, artificial intelligence will act less as a replacement than as a co-scientist, suggesting and refining experiments, exploring alternative hypotheses, and supporting interpretation, while experimental biology continues to provide the indispensable anchor for discovery.

The trajectory is unmistakable. In the coming years, we are likely to witness the first true AI-driven breakthroughs. In hindsight, these may seem obvious, but only because a new lens will have made them visible. The combination of artificial intelligence and multimodal cell biology shows that the secrets of life yield not only to those who look closely but also to those who can synthesize across angles, perspectives, levels of organization, and scales. By bringing computation into the very core of experimentation, we are breathing life into data, creating models that learn alongside our understanding. The result is a more nuanced grasp of how cells cooperate to create organisms and how we might wisely guide them, whether in treating disease or engineering biology. A new era of cellular insight is dawning and its defining feature is that artificial intelligence is not just studying biology, it is becoming part of it – an active participant in humanity’s quest to know itself.

ABOUT THE AUTHORS

Charlotte Bunne is Assistant Professor in Computer and Communication Sciences and Life Sciences at the École Polytechnique Fédérale de Lausanne (EPFL). She is also part of the Swiss Institute for Experimental Cancer Research (ISREC) and the EPFL AI Center. She has published in such journals as *Nature Methods*, *Cell*, and *Bioinformatics*.

Aviv Regev, a Member of the American Academy since 2022, is Head and Executive Vice President of Genentech Research and Early Development (gRED) and founding Co-Chair of the Human Cell Atlas. She has also served as Chair of the Faculty and Core Member at the Broad Institute of MIT and Harvard, Professor of Biology at MIT, and as a Howard Hughes Medical Institute Investigator. She is the author or coauthor of over four hundred peer-reviewed articles and has recently published in such journals as *Nature*, *Science*, and *Cell*.

ENDNOTES

- ¹ Evan Z. Macosko, Anindita Basu, Rahul Satija, et al., “Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets,” *Cell* 161 (5) (2015): 1202–1214; Allon M. Klein, Linas Mazutis, Ilke Akartuna, et al., “Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells,” *Cell* 161 (5) (2015): 1187–1201; Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, et al., “Massively Parallel Digital

- Transcriptional Profiling of Single Cells,” *Nature Communications* 8 (2017): 14049; Charlotte Giesen, Hao A. O. Wang, Denis Schapiro, et al., “Highly Multiplexed Imaging of Tumor Tissues with Subcellular Resolution by Mass Cytometry,” *Nature Methods* 11 (4) (2014): 417–422; and Marlon Stoeckius, Christoph Hafemeister, William Stephenson, et al., “Simultaneous Epitope and Transcriptome Measurement in Single Cells,” *Nature Methods* 14 (9) (2017): 865–868.
- ² Jellert T. Gaublomme, Nir Yosef, Youjin Lee, et al., “Single-Cell Genomics Unveils Critical Regulators of Th17 Cell Pathogenicity,” *Cell* 163 (6) (2015): 1400–1412; and Dominic Grün, Anna Lyubimova, Lennart Kester, et al., “Single-Cell Messenger RNA Sequencing Reveals Rare Intestinal Cell Types,” *Nature* 525 (7568) (2015): 251–255.
- ³ Aviv Regev, Sarah A. Teichmann, Eric S. Lander, et al., “The Human Cell Atlas,” *eLife* 6 (2017), <https://doi.org/10.7554/eLife.27041>.
- ⁴ Ibid.
- ⁵ Atray Dixit, Oren Parnas, Biyu Liu, et al., “Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens,” *Cell* 167 (7) (2016): 1853–1866; Britt Adamson, Thomas M. Norman, Marco Jost, et al., “A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response,” *Cell* 167 (7) (2016): 1867–1882; and Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, et al., “Mapping Information-Rich Genotype-Phenotype Landscapes with Genome-Scale Perturb-Seq,” *Cell* 185 (14) (2022): 2559–2575.
- ⁶ Thomas M. Norman, Max A. Horlbeck, Joseph M. Replogle, et al., “Exploring Genetic Interaction Manifolds Constructed from Rich Single-Cell Phenotypes,” *Science* 365 (6455) (2019): 786–793; and Jennifer E. Rood, Anna Hupalowska, and Aviv Regev, “Toward a Foundation Model of Causal Cell and Tissue Biology with a Perturbation Cell and Tissue Atlas,” *Cell* 187 (17) (2024): 4520–4545.
- ⁷ Jennifer E. Rood, Aidan Maartens, Anna Hupalowska, et al., “Impact of the Human Cell Atlas on Medicine,” *Nature Medicine* 28 (12) (2022): 2486–2496.
- ⁸ Kimberly Siletti, Rebecca Hodge, Alejandro Mossi Albiach, et al., “Transcriptomic Diversity of Cell Types Across the Adult Human Brain,” *Science* 382 (6667) (2023): 175.
- ⁹ Daniel T. Montoro, Adam L. Haber, Moshe Biton, et al., “A Revised Airway Epithelial Hierarchy Includes CFTR-Expressing Ionocytes,” *Nature* 560 (7718) (2018): 319–324; Lindsey W. Plasschaert, Rapolas Žilionis, Rayman Choo-Wing, et al., “A Single-Cell Atlas of the Airway Epithelium Reveals the CFTR-Rich Pulmonary Ionocyte,” *Nature* 560 (7718) (2018): 377–381; and Eugene Drokhlyansky, Christopher S. Smillie, Nicholas Van Wittenberghe, et al., “The Human and Mouse Enteric Nervous System at Single-Cell Resolution,” *Cell* 182 (6) (2020): 1606–1622.
- ¹⁰ Rood, Maartens, Hupalowska, et al., “Impact of the Human Cell Atlas on Medicine.”
- ¹¹ Sai Ma, Bing Zhang, Lindsay M. LaFave, et al., “Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin,” *Cell* 183 (4) (2020): 1103–1116; and Stoeckius, Hafemeister, Stephenson, et al., “Simultaneous Epitope and Transcriptome Measurement in Single Cells.”
- ¹² Ibid.
- ¹³ Hari Shroff, Ilaria Testa, Florian Jug, and Suliana Manley, “Live-Cell Imaging Powered by Computation,” *Nature Reviews Molecular Cell Biology* 25 (6) (2024): 443–463; Gioele La Manno, Ruslan Soldatov, Amit Zeisel, et al., “RNA Velocity of Single Cells,” *Nature* 560 (7719) (2018):

- 494–498; Laleh Haghverdi, Maren Büttner, F. Alexander Wolf, et al., “Diffusion Pseudotime Robustly Reconstructs Lineage Branching,” *Nature Methods* 13 (10) (2016): 845–848; Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, et al., “Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming,” *Cell* 176 (6) (2019): 928–943; and Dominik Klein, Giovanni Palla, Marius Lange, et al., “Mapping Cells Through Time and Space with Moscot,” *Nature* 638 (8052) (2025): 1065–1075.
- ¹⁴ Joshua A. Weinstein, Aviv Regev, and Feng Zhang, “DNA Microscopy: Optics-Free Spatio-Genetic Imaging by a Stand-Alone Chemical Reaction,” *Cell* 178 (1) (2019): 229–241; Patrik L. Ståhl, Fredrik Salmén, Sanja Vickovic, et al., “Visualization and Analysis of Gene Expression in Tissue Sections by Spatial Transcriptomics,” *Science* 353 (6294) (2016): 78–82; Samuel G. Rodriques, Robert R. Stickels, Aleksandrina Goeva, et al., “Slide-Seq: A Scalable Technology for Measuring Genome-Wide Expression at High Spatial Resolution,” *Science* 363 (6434) (2019): 1463–1467; and Nianchao Qian and Joshua A. Weinstein, “Spatial Transcriptomic Imaging of an Intact Organism Using Volumetric DNA Microscopy,” *Nature Biotechnology*, March 27, 2025, <https://doi.org/10.1038/s41587-025-02613-z>.
- ¹⁵ Kok Hao Chen, Alistair N. Boettiger, Jeffrey R. Moffitt, et al., “Spatially Resolved, Highly Multiplexed RNA Profiling in Single Cells,” *Science* 348 (6233) (2015): 1360–1363; Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, et al., “Transcriptome-Scale Super-Resolved Imaging in Tissues by RNA seqFISH+,” *Nature* 568 (7751) (2019): 235–239; Giesen, Wang, Schapiro, et al., “Highly Multiplexed Imaging of Tumor Tissues with Subcellular Resolution by Mass Cytometry”; and Yuri Goltsev, Nikolay Samusik, Julia Kennedy-Darling, et al., “Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging,” *Cell* 174 (4) (2018): 968–981.
- ¹⁶ Regev, Teichmann, Lander, et al., “Science Forum: The Human Cell Atlas”; Rood, Maartens, Hupalowska, et al., “Impact of the Human Cell Atlas on Medicine”; and Rood, Hupalowska, and Regev, “Toward a Foundation Model of Causal Cell and Tissue Biology with a Perturbation Cell and Tissue Atlas.”
- ¹⁷ Macosko, Basu, Satija, et al., “Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets.”
- ¹⁸ Ibid; and Romain Lopez, Jeffrey Regier, Michael B. Cole, et al., “Deep Generative Modeling for Single-Cell Transcriptomics,” *Nature Methods* 15 (12) (2018): 1053–1058.
- ¹⁹ Andrew Butler, Paul Hoffman, Peter Smibert, et al., “Integrating Single-Cell Transcriptomic Data Across Different Conditions, Technologies, and Species,” *Nature Biotechnology* 36 (5) (2018): 411–420; Tim Stuart, Andrew Butler, Paul Hoffman, et al., “Comprehensive Integration of Single-Cell Data,” *Cell* 177 (7) (2019): 1888–1902; Brian Hie, Bryan Bryson, and Bonnie Berger, “Efficient Integration of Heterogeneous Single-Cell Transcriptomes Using Scanorama,” *Nature Biotechnology* 37 (6) (2019): 685–691; Ilya Korsunsky, Nghia Milard, Jean Fan, et al., “Fast, Sensitive and Accurate Integration of Single-Cell Data with Harmony,” *Nature Methods* 16 (12) (2019): 1289–1296; Charles Comiter, Eeshit Dhaval Vaishnav, Metamia Ciampricotti, et al., “Inference of Single Cell Profiles from Histology Stains with the Single-Cell Omics from Histology Analysis Framework (SCHAF),” *bioRxiv* (2023), <https://www.biorxiv.org/content/10.1101/2023.03.21.533680v1>; and Weiqing Chen, Pengzhi Zhang, Tu N. Tran, et al., “A Visual-Omics Foundation Model to Bridge Histopathology with Spatial Transcriptomics,” *Nature Methods* 22 (7) (2025): 1568–1582.
- ²⁰ Charlotte Bunne, Stefan G. Stark, Gabriele Gut, et al., “Learning Single-Cell Perturbation Responses Using Neural Optimal Transport,” *Nature Methods* 20 (11) (2023): 1759–1768; and Yusuf Roohani, Kexin Huang, and Jure Leskovec, “Predicting Transcriptional

- Outcomes of Novel Multigene Perturbations with GEARS,” *Nature Biotechnology* 42 (6) (2024): 927–935.
- ²¹ Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al., “On the Opportunities and Risks of Foundation Models,” arXiv (2021), <https://arxiv.org/abs/2108.07258>.
- ²² Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Association for Computational Linguistics, 2019), 4171–4186, <https://aclanthology.org/events/naacl-2019/#n19-1>; and Alec Radford, Jeffrey Wu, Rewon Child, et al., “Better Language Models and Their Implications,” OpenAI News, February 14, 2019, <https://openai.com/index/better-language-models>.
- ²³ Rood, Maartens, Hupalowska, et al., “Impact of the Human Cell Atlas on Medicine”; Haotian Cui, Chloe Wang, Hassaan Maan, et al., “scGPT: Toward Building a Foundation Model for Single-Cell Multi-Omics Using Generative AI,” *Nature Methods* 21 (8) (2024): 1470–1480; Minsheng Hao, Jing Gong, Xin Zeng, et al., “Large-Scale Foundation Model on Single-Cell Transcriptomics,” *Nature Methods* 21 (8) (2024): 1481–1491; Xi Fu, Shentong Mo, Alejandro Buendia, et al., “A Foundation Model of Transcription Across Human Cell Types,” *Nature* 637 (8047) (2025): 965–973; and Graham Heimberg, Tony Kuo, Daryle J. DePianto, et al., “A Cell Atlas Foundation Model for Scalable Search of Similar Human Cells,” *Nature* 638 (8052) (2025): 1085–1094.
- ²⁴ Cui, Wang, Maan, et al., “scGPT: Toward Building a Foundation Model for Single-Cell Multi-Omics Using Generative AI”; Hao, Gong, Zeng, et al., “Large-Scale Foundation Model on Single-Cell Transcriptomics”; Fu, Mo, Buendia, et al., “A Foundation Model of Transcription Across Human Cell Types”; Heimberg, Kuo, DePianto, et al., “A Cell Atlas Foundation Model for Scalable Search of Similar Human Cells”; and Christina V. Theodoris, Ling Xiao, Anant Chopra, et al., “Transfer Learning Enables Predictions in Network Biology,” *Nature* 618 (7965) (2023): 616–624.
- ²⁵ Ashish Vaswani, Noam Shazeer, Niki Parmar, et al., “Attention Is All You Need,” in *Advances in Neural Information Processing Systems* 30 (NIPS 2017), ed. Isabelle Guyon, Ulrike Von Luxburg, Samy Bengio, et al. (Neural Information Processing Systems Foundation, 2017).
- ²⁶ Cui, Wang, Maan, et al., “scGPT: Toward Building a Foundation Model for Single-Cell Multi-Omics Using Generative AI”; and Johann Wenckstern, Eeshaan Jain, Yexiang Cheng, et al., “AI-Powered Virtual Tissues from Spatial Proteomics for Clinical Diagnostics and Biomedical Discovery,” arXiv (2025), <https://doi.org/10.48550/arXiv.2501.06039>.
- ²⁷ Leah V. Schaffer, Mengzhou Hu, Gege Qian, et al., “Multimodal Cell Maps as a Foundation for Structural and Functional Genomics,” *Nature* 642 (8066) (2025): 222–231.
- ²⁸ Comiter, Vaishnav, Ciampricotti, et al., “Inference of Single Cell Profiles from Histology Stains with the Single-Cell Omics from Histology Analysis Framework (SCHAF)” ; and Chen, Zhang, Tran, et al., “A Visual-Omics Foundation Model to Bridge Histopathology with Spatial Transcriptomics.”
- ²⁹ Richard J. Chen, Tong Ding, Ming Y. Lu, et al., “Towards a General-Purpose Foundation Model for Computational Pathology,” *Nature Medicine* 30 (3) (2024): 850–862; Thomas Hayes, Roshan Rao, Halil Akin, et al., “Simulating 500 Million Years of Evolution with a Language Model,” *Science* 387 (6736) (2025): 850–858; and Eric Nguyen, Michael Poli, Matthew G. Durrant, et al., “Sequence Modeling and Design from Molecular to Genome Scale with Evo,” *Science* 386 (6723) (2024), <https://doi.org/10.1126/science.ad09336>.

- ³⁰ Cui, Wang, Maan, et al., “scGPT: Toward Building a Foundation Model for Single-Cell Multi-Omics Using Generative AI”; and Theodoris, Xiao, Chopra, et al., “Transfer Learning Enables Predictions in Network Biology.”
- ³¹ Heimberg, Kuo, DePianto, et al., “A Cell Atlas Foundation Model for Scalable Search of Similar Human Cells”; and Wenckstern, Jain, Cheng, et al., “AI-Powered Virtual Tissues from Spatial Proteomics for Clinical Diagnostics and Biomedical Discovery.”
- ³² Cui, Wang, Maan, et al., “scGPT: Toward Building a Foundation Model for Single-Cell Multi-Omics Using Generative AI”; Hao, Gong, Zeng, et al., “Large-Scale Foundation Model on Single-Cell Transcriptomics”; and Fu, Mo, Buendia, et al., “A Foundation Model of Transcription Across Human Cell Types.”
- ³³ Heimberg, Kuo, DePianto, et al., “A Cell Atlas Foundation Model for Scalable Search of Similar Human Cells.”
- ³⁴ Chen, Zhang, Tran, et al., “A Visual-Omics Foundation Model to Bridge Histopathology with Spatial Transcriptomics.”
- ³⁵ Dixit, Parnas, Liu, et al., “Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens”; Adamson, Norman, Jost, et al., “A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response”; and Replogle, Saunders, Pogson, et al., “Mapping Information-Rich Genotype-Phenotype Landscapes with Genome-Scale Perturb-Seq.”
- ³⁶ Theodoris, Xiao, Chopra, et al., “Transfer Learning Enables Predictions in Network Biology.”
- ³⁷ Bunne, Stark, Gut, et al., “Learning Single-Cell Perturbation Responses Using Neural Optimal Transport”; Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis, “scGen Predicts Single-Cell Perturbation Responses,” *Nature Methods* 16 (8) (2019): 715–721; and Roohani, Huang, and Leskovec, “Predicting Transcriptional Outcomes of Novel Multi-gene Perturbations with GEARS.”
- ³⁸ Kexin Huang, Serena Zhang, Hanchen Wang, et al., “Biomni: A General-Purpose Biomedical AI Agent,” bioRxiv (2025), <https://www.biorxiv.org/content/10.1101/2025.05.30.656746v1>; and Eric Wang, Samuel Schmidgall, Paul F. Jaeger, et al., “TxGemma: Efficient and Agentic LLMs for Therapeutics,” arXiv (2025), <https://doi.org/10.48550/arXiv.2504.06196>.
- ³⁹ Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, et al., “Generative Adversarial Nets,” in *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, ed. Zoubin Ghahramani, Max Welling, Corinna Cortes, et al. (Neural Information Processing Systems Foundation, 2014); Jonathan Ho, Ajay N. Jain, and Pieter Abbeel, “Denoising Diffusion Probabilistic Models,” in *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, ed. Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, et al. (Neural Information Processing Systems Foundation, 2020); and Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, et al., “Flow Matching for Generative Modeling,” paper presented at the Eleventh International Conference on Learning Representations, Kigali, Rwanda, May 1–5, 2023, <https://openreview.net/pdf?id=PqvMRDCJT9t>.
- ⁴⁰ Eeshit Dhaval Vaishnav, Carl G. de Boer, Jennifer Molinet, et al., “The Evolution, Evolvability and Engineering of Gene Regulatory DNA,” *Nature* 603 (7901) (2022): 455–463.
- ⁴¹ Richard S. Sutton and Andrew G. Barto, *Reinforcement Learning: An Introduction*, 1st ed. (MIT Press, 1998).

- ⁴² Charlotte Bunne, Yusuf Roohani, Yanay Rosen, et al., “How to Build the Virtual Cell with Artificial Intelligence: Priorities and Opportunities,” *Cell* 187 (25) (2024): 7045–7063.
- ⁴³ Cui, Wang, Maan, et al., “scGPT: Toward Building a Foundation Model for Single-Cell Multi-Omics Using Generative AI”; Hao, Gong, Zeng, et al., “Large-Scale Foundation Model on Single-Cell Transcriptomics”; Fu, Mo, Buendia, et al., “A Foundation Model of Transcription Across Human Cell Types”; and Heimberg, Kuo, DePianto, et al., “A Cell Atlas Foundation Model for Scalable Search of Similar Human Cells.”
- ⁴⁴ Huang, Zhang, Wang, et al., “Biomni: A General-Purpose Biomedical AI Agent”; and Wang, Schmidgall, Jaeger, et al., “TxGemma: Efficient and Agentic LLMs for Therapeutics.”
- ⁴⁵ Bunne, Roohani, Rosen, et al., “How to Build the Virtual Cell with Artificial Intelligence: Priorities and Opportunities.”
- ⁴⁶ Chen, Zhang, Tran, et al., “A Visual-Omics Foundation Model to Bridge Histopathology with Spatial Transcriptomics”; and Ståhl, Salmén, Vickovic, et al., “Visualization and Analysis of Gene Expression in Tissue Sections by Spatial Transcriptomics.”
- ⁴⁷ Bunne, Stark, Gut, et al., “Learning Single-Cell Perturbation Responses Using Neural Optimal Transport”; Roohani, Huang, and Leskovec, “Predicting Transcriptional Outcomes of Novel Multigene Perturbations with GEARS”; and Huang, Zhang, Wang, et al., “Biomni: A General-Purpose Biomedical AI Agent.”

Building the Drug Discovery Engine of the Future with AI-Empowered Nodal Biology

Anna Greka

Inspired by the visionary predictions of Jules Verne, this essay proposes that integrating artificial intelligence with a paradigm called “nodal biology” will revolutionize the discovery of treatments and cures. Current drug development is slow, costly, and failure-prone, constrained by the challenge of drug target identification. Solving the “cell perturbation prediction problem” – predicting how human cells respond to any disease-causing perturbation – is key to accelerating successful drug target identification. Nodal biology, the discovery of shared druggable mechanisms (nodes) among seemingly disparate diseases, offers a scalable approach to generate the high-quality data needed to train cell prediction AI models. As an example, a cargo receptor node was identified, linking dozens of genetic diseases and leading to a new drug candidate. The synergistic combination of human scientific intuition and AI-empowered nodal biology is essential for building the biomedical innovation engine of the future, ultimately accelerating treatments for all human diseases.

Writing more than 150 years ago, Jules Verne was celebrated for his adventure novels, incredible stories that integrated science fiction with explorations of uncharted territories. Many of his predictions ultimately materialized. He imagined submarines diving to the depths of the oceans and rockets taking humans to space. Quite remarkably, in *The Begum's Fortune* (1879), Verne incorporated biology into his writing. In this lesser-known novel, Verne explores the concept of artificial life and the future of biology. His characters endeavor to create a living organism through the manipulation of biological processes, thereby anticipating ideas later explored in genetic engineering. Verne's intuition and imagination allowed him to make visionary predictions, well beyond the scientific knowledge of his time.

As we now stand at the precipice of a new era for humanity empowered by artificial intelligence, can we – like a modern-day Jules Verne – envision a future when we can cure all human diseases? Can we imagine fully understanding the functional architecture of all human cells, then organs, and ultimately the entire body in health and disease? And can this knowledge enable effective methods for the prediction, control, and optimization of their inner workings? Success in this realm will bring

a revolution to biomedicine: the ability to predict and control the response of any human cell to any disease-related perturbation will promote our diagnostic capabilities (detect disruptions in specific cells with precision) and, more important, our capabilities to treat diseases in an exquisitely targeted fashion (direct drugs to specific cell disruptions), ultimately extending healthy lifespan. Is this possible or is it science fiction? And if it is possible, what will it take to get there?

Recent advances in AI have propelled us into a golden era in biomedicine. This is exemplified by the groundbreaking work of Demis Hassabis, John Jumper, and David Baker to use AI-empowered models to predict the structure of proteins and to learn to create entirely new proteins, solving the “protein structure prediction problem,” for which they were awarded the 2024 Nobel Prize in Chemistry. Armed with these tools, the scientific community can now pursue the development of therapies with greater speed and efficiency. Furthermore, AI-driven automation systems and advanced imaging approaches that use computer vision and related deep-learning models combined with new capabilities in genetic engineering collectively empower scientists to pursue biological experiments at unprecedented scale. In this essay, I delve into how AI is transforming biomedicine, the inherent challenges, and the promising future trajectories. This perspective is shaped by my own research as a physician-scientist and cell biologist, and I consider how the combination of a new paradigm called *nodal biology* can be fully harnessed with AI-enabled tools to revolutionize the specificity, speed, and efficiency with which we develop much needed new treatments.

Building innovative therapies is difficult. To this day, the inherent challenges and risks associated with the development of novel, targeted therapies remain substantial. As of 2026, the probability of a new drug progressing from initial human studies to regulatory approval remains at a mere 13 percent, a rate that has not changed in decades. Of the drugs that are approved, the average cost of research and development (including clinical trials) is more than \$1 billion each. And the timelines to approval are very long: it often takes more than ten years for a treatment to be approved and become available to patients. Additional factors, such as shifts in regulatory policies and geopolitical considerations, exert profound effects on our ability to develop new therapies. Nevertheless, improving on the low success rate in research and development continues to be the core challenge to building a productive innovation engine.

For all practical purposes, the innovation challenge can be broken into three simple but critical components:

1. Target identification (ID) – finding the right drug targets with therapeutic significance for as many diseases as possible;

2. Optimal drug development – finding the right therapeutic modality for each target; and
3. Clinical deployment – giving the right therapy to a well-characterized patient population with clearly defined clinical readouts (that is, “endpoints”) to maximize therapeutic benefit.

AI is already proving massively catalytic for components two and three, with astonishing progress taking place in real time. Regarding component two – optimal drug development – AI models that have solved the protein structure prediction problem such as AlphaFold are turbocharging rational drug design.¹ This means that we can now monitor and control the process of designing the right drug to fit into the right “pocket” within a specific, well-defined molecular target, and we can do it by visualizing the drug bound to its target in 3D on our computer screens (*in silico*). The time it takes to build new drugs (such as small molecules or therapeutic antibodies) is thereby being markedly shortened and made more efficient every day. Regarding component three – clinical deployment – AI is accelerating patient clinical data aggregation, biomarker identification, and patient stratification, which should empower more-successful clinical trials in years to come. Despite ongoing challenges with data aggregation and harmonization, coupled with scientific, legal, and ethical barriers, AI is likely to facilitate more efficiently run clinical trials in the years ahead. Components two and three are already benefiting from the tools of AI. While critical improvements are still needed and will undoubtedly continue, these two components are not what currently constrains our ability to innovate.

The most challenging part of our innovation engine continues to be component one, target ID: the fundamental work required to identify molecular targets in human cells that can be perturbed for therapeutic benefit (a given molecular target can be inhibited or activated). I strongly believe that addressing the foundational target ID component of the innovation challenge will require solving the “cell perturbation prediction problem”: that is, understanding and predicting the repertoire of responses by human cells to any given disease-causing perturbation. Solving this challenge is likely to unleash an innovation engine ultimately capable of addressing human diseases at scale.

As a physician-scientist, I have training in both medicine and scientific research, and my core objective is to translate fundamental laboratory discoveries into therapeutic strategies that can benefit patients: that is, to treat all diseases for all patients, with utmost expediency. Given the complexities and difficulties previously outlined, how can we realistically aspire to reach this goal? I am optimistic that progress will come from creatively combining our finest current and future AI tools with human intelligence, or what I call “scientific intuition.” Marshaling our resources to solve the cell perturbation prediction problem will bring us closer to achieving the lofty goal of treating all human diseases.

The key bottleneck in solving this foundational problem is obtaining cell perturbation data of large enough magnitude and quality to train cell-prediction AI models. Deep-learning models require data of a scale that is unprecedented in cell biology. My current work is focused on generating such key datasets by starting with human genetic diseases, specifically monogenic disorders in which a mutation in the coding sequence of a given gene leads to a defined disease phenotype (for example, cystic fibrosis, a debilitating lung disease caused by specific genetic defects in a gene called *CFTR*). We have selected to begin with monogenic disorders for which the genetic cause is known because we can readily introduce these human pathogenic mutations into any cell of our choice to develop a faithful disease model using genetic engineering approaches like base editing.² Modeling thousands of human pathogenic mutations (introducing thousands of genetic perturbations) in diverse human cell types (neurons, cardiac myocytes, liver cells, and so on) and interrogating them with robust scalable readouts (single cell transcriptomics or proteomics, reading the repertoire of mRNA messages or proteins found in each of these cells) will generate a high-quality dataset with rigorous controls. In other words, this approach will generate data of the magnitude and quality needed to train AI models that can help us solve the cell perturbation prediction problem.

Addressing the cell perturbation prediction problem will be greatly facilitated by *nodal biology*. Over the last few years, my team and I have discovered that many clinically unrelated genetic disorders share common – most often uncharacterized – mechanisms, which we call *nodal pathways*. In other words, seemingly disparate diseases (for example, an inherited form of blindness, a familial form of Alzheimer’s disease, and a genetic kidney disease) ultimately converge on the same node, representing a shared druggable target (for example, TMED cargo receptors).³ We thus introduced the concept of *nodal biology*: the discovery of nonobvious, previously unseen connections between seemingly distinct diseases. This paradigm allows us to design large-scale interrogations of all monogenic diseases in a massively parallel manner to identify the full complement of biological nodes that connect subsets of these diseases in previously unseen nodal clusters.

How many monogenic diseases are there? With advances in DNA sequencing technologies, we can now readily detect mutations in more than eight thousand (out of a total of twenty thousand) human genes that drive these devastating chronic conditions; however, even as the pool of known mutations is growing, fewer than 5 percent of them have an available treatment. Genetic diseases collectively affect thirty million people across the United States, most arising during childhood, and more than four hundred million people globally.⁴ A key reason for the significant lack of therapies is that most research on genetic diseases has remained siloed

among small teams of scientists painstakingly working on one gene/one disease at a time. In other words, the fundamental challenges associated with the “one-by-one genetic defect” approach are: 1) scale – work happens in individual laboratories and few discoveries are able to reach critical mass in terms of validation and downstream clinical impact; 2) investability – the patient populations are fragmented into small subsets that do not attract sufficient biopharmaceutical interest; and 3) biological convergence – the one-by-one approach fails to leverage recent insights that different diseases converge on the same biological pathways. Nodal biology can address these challenges.

To be clear, the concept of convergent disease mechanisms is not new. In some cases, taking a known drug with a well-defined target from one disease into several others that share the same mechanism (what is sometimes called “indication extension”) is relatively obvious. For example, such an approach targeting the alternative complement pathway – an immune mechanism designed to destroy invading human pathogens – is an approved therapy that is helping patients with chronic and devastating kidney, blood, neuromuscular, and eye diseases.⁵ Where nodal biology truly shines is in uncovering nonobvious, previously unseen connections between seemingly different diseases. Our team discovered such a biological node, a point of convergence on a single druggable target at which mechanistic connections were identified between dozens of clinically disparate but connected genetic diseases affecting the eye, the kidney, the brain, and more.⁶ Given that potentially hundreds more nodes underlie approximately eight thousand human genetic diseases caused by up to one hundred thousand different mutations, bold innovation is crucial. To pave the path to much needed nodal therapies, we must revolutionize our ability to identify nodal pathways and druggable targets by deploying cell perturbation studies at scale. A bold vision for the future is to uncover as many druggable nodes as possible in a massively parallel fashion.

Gaining insights from nodal biology will require the use of scalable tools, including: genome editing techniques that can allow us to introduce disease-causing mutations in human cells at scale and thus to generate cellular models of thousands of different genetic diseases in parallel; high-throughput methods of reading the effects of thousands of genetic perturbations into a collection of human cells; and the ability to turn up or down any gene of interest in parallel in a collection of human cells.⁷ These tools are readily available today and can be scaled to serve the needs of any nodal biology project.

Nodal biology will also require computational methodologies (many still in development) to collect and analyze unprecedented amounts of human cell perturbation data to identify convergent nodes.⁸ We can subsequently use these data to train cell-perturbation AI models in an iterative fashion, bringing us closer to solving the cell prediction problem: that is, the ability to *predict* the response of human cells to a given perturbation that was not included in the training data – all *in silico*.

How would nodal biology solving the cell prediction problem – thereby securing component one of the innovation engine of the future – enhance our drug target discovery process? Imagine that scientists have uncovered multiple candidates as the potential drug target that could cure a devastating genetic disease. Without having to do expensive and time-consuming studies that could take years and millions of dollars, can they use our nodal biology–derived cell model to predict with confidence which of their candidates is the best drug target? Can they predict if turning the target up or down will be required for therapeutic benefit? And can they predict if drugging this target will be toxic to other cells or organs in the human body? Solving the cell prediction problem and building accessible models for the scientific community – much like AlphaFold for predicting protein structures – will vastly accelerate our innovation engine by empowering scientists to test multiple potential targets *in silico* with high confidence, ease, and efficiency.

To illustrate the arc of discovery in nodal biology, I will share an example from my own work. This discovery exemplifies how scientific intuition – derived from deep knowledge of cell biology and years of working in the laboratory – led us to a nodal target; and how combining this insight with AI tools like AlphaFold drove the discovery of a drug candidate for the treatment of several devastating diseases.

Several years ago, our team was tasked with solving a perplexing medical mystery. As is often the case with such inquiries, the stakes were considerable: the mystery involved a large family that has been suffering with kidney failure and untimely death for many generations. What, then, was the underlying cause of their affliction? Through concerted efforts, the etiology of this disease was finally revealed to be a mutation, specifically a single cytosine (C) misspelling in the DNA. This singular genetic error is sufficient to induce severe kidney damage in these patients.⁹

As a brief reminder, DNA is the genetic code, comprising three billion letters of information meticulously packaged within the nuclei of each of approximately thirty trillion cells in our bodies. Within cells, DNA is transcribed into RNA, and RNA, in turn, is translated into proteins, which are the functional units employed by our cells to execute their intricate functions.

Tragically, in the case of the family in question, the solitary DNA misspelling leads to the production of a malformed, dysfunctional, and toxic protein that progressively accumulates within their cells, ultimately resulting in cell death and kidney failure. In the absence of cell-prediction models (that is, in the absence of component-one solutions), we had no choice but to pursue answers by what I call “molecular sleuthing,” otherwise defined as the judicious integration of our most advanced experimental tools with scientific intuition. After a few years of such investigations, we finally found that this toxic protein accumulates within cells due to the involvement of another molecule, a “cargo receptor.”¹⁰ These cargo recep-

tors sequester the toxic protein and become ensnared in a cellular “traffic jam.” Over time, the toxic cargo accrues, ultimately leading to cell death. In essence, after more than five years of hard work, we found that the cargo receptors were the right (therapeutically relevant) drug target for this disease.

Based on these insights, our objective was to identify a drug capable of interfering with the cargo receptors and directing them to release their toxic cargo toward the cell’s waste disposal facility, known as the lysosome. We embarked on the search for a suitable drug candidate. We had some initial hints that a compound could be effective in removing toxic cargo, but that candidate proved relatively ineffective and raised concerns about potential toxicity to brain cells (otherwise known as “off-target” toxicity). Fortunately, by this time in the project, we were armed with AlphaFold, a powerful tool for drug discovery (and an excellent solution for component two of our innovation engine). Using AlphaFold, we were able to visualize the precise arrangement of each atom composing the cargo receptor target. This capability, in turn, facilitated the rational design of small molecules that can fit precisely into the exact right pocket in the cargo receptor target to achieve the desired specific effect: in this instance, to direct the cargo receptors and their toxic cargo toward the lysosome. Consequently, within an unusually short timeframe, we identified a new drug candidate that fits precisely into the designated pocket of this cargo receptor. When tested in human cells and mice, it performed exactly as intended: it steered the toxic cargo toward the lysosome, thereby clearing the cells of the toxic traffic jam. It is satisfying to know that after several years of work, this discovery is now advancing toward clinical trials in patients – and the family that inspired all this work is now our close partner in driving this program to the clinic.

Our investigations did not stop there. We subsequently discovered that the same drug target, the cargo receptors, are implicated in many different genetic diseases that involve the accumulation of different toxic proteins in different organs. These are severe disorders such as blindness, liver failure, and Alzheimer’s disease. In total, we estimate that more than fifty diseases, involving millions of patients, could ultimately be addressed by targeting the same cargo receptor system.

In essence, we discovered a nonobvious “node” that illuminated previously unknown connections between multiple unrelated genetic diseases, which in turn led to the realization that nodal biology could hold the key to accelerating the discovery of drug targets for the treatment of heretofore incurable diseases. In this new paradigm, we defined a node as a point of convergence on a single druggable target, enabled by the identification of previously unrecognized, shared biological pathways among seemingly disparate but now newly interconnected diseases. This work led us to ask the next fundamental question: Can we identify all the nodes that underpin all human genetic diseases? This forms the foundation for our current and future work on AI-empowered nodal biology.

Beyond monogenic diseases, solving the cell prediction problem requires us to perturb human cells with molecules and factors that drive devastating disorders such as obesity, cancer, autoimmune conditions, mental health conditions, and diseases of aging. We also need to scale our investigations to interrogating billions of human cells in parallel with readouts that capture disease-specific changes in cellular mRNA, proteins, and metabolites. Furthermore, progress in this grand challenge in biomedicine will require a shift to a new biomedical research model that incorporates AI-enabled robots and automation contributing alongside AI-empowered scientists. In the final analysis, the synergistic combination of scientific intuition and current and future AI tools can propel an unprecedented acceleration of our innovation engine toward new targeted therapies. I am confident that the next few years will prove catalytic, as we progress from identifying a few druggable nodes to progressively *predicting* which nodes to target until we can readily treat all human diseases with speed and precision.

As we pursue the merger of nodal biology with AI to build the drug discovery innovation engine of the future, it is worth noting that the intrinsic complexity of biological data poses significant challenges. Rigorous validation processes and iterative feedback loops are critical to ensure the reliability and accuracy of AI-driven insights. A key limitation lies in the ability to train deep-learning AI models with data of consistently high quality. Biological data often present with inherent variability due to experimental conditions, individual differences, and technological limitations, making it difficult to acquire sufficiently comprehensive datasets for robust model training. For example, a cellular perturbation dataset in Cambridge, Massachusetts, may not be identical to a dataset produced with the same protocol in Cambridge, England. The presence of noise, missing values, and inconsistencies within datasets further exacerbates the challenge, requiring advanced preprocessing techniques to clean and prepare the data for optimal model performance. Another potential consideration is the challenge of generating a massive dataset at a single lab/location with highly rigorous procedures before the data can be widely shared to train cell perturbation models. Without addressing these fundamental issues of biological data complexity and quality, the full potential of deploying cell-prediction AI models may remain constrained.

Nodal biology anchored on genetic diseases is limited in that it does not address a critical driver for many other human disorders: namely, the fact that disease processes also occur in the cross talk between different cell types, such as between immune cells interacting with brain cells or liver cells or heart cells. Developing the tools to measure perturbations in cell-cell interactions and more complex systems such as intact organs may ultimately be needed before we can train models that can give us predictive power over complex processes such as cancer metastasis or the interactions underlying schizophrenia. Our innovation engine will benefit from an exploration of systems of higher complexity once our nodal biology

paradigm for genetic diseases has given us the foundational tools to take on new challenges.

Advances in any complex scientific field, particularly one as multifaceted as the intersection of AI and biomedicine, require a deeply integrated, interdisciplinary approach. To achieve breakthroughs in building the innovation engine of the future, it is paramount to train a workforce that integrates expertise in AI/machine learning, biology, and medicine. Currently, a significant hurdle lies in the cross talk between these disciplines: each often operates with its own specialized lexicon and conceptual frameworks. Ongoing and future programs in our laboratories and universities should therefore focus on enhancing the opportunities for deep collaborative work between AI specialists and biomedical scientists.

AI models are only as good as the data they are trained on and the biological relevance of the questions they are designed to address. Biologists and physicians possess domain knowledge, understanding the intricacies of biological systems, experimental design, and the clinical implications of research findings. AI specialists offer the computational tools and analytical techniques to process vast datasets and build predictive models. The challenge lies in synthesizing these strengths, ensuring that AI algorithms are developed with a deep understanding of biology and human disease and that biological questions are framed in a way that can be effectively tackled by AI.

In Douglas Adams's *The Hitchhiker's Guide to the Galaxy*, the supercomputer Deep Thought, after eons of computation, reveals the answer to the "Ultimate Question of Life, the Universe, and Everything": it is "42." This answer was a huge disappointment to those awaiting a response – but they had not asked "the right question." The true challenge in AI-empowered biomedical innovation lies not in obtaining an answer, but in identifying and formulating the relevant question itself. This narrative serves as a poignant and cautionary tale for our engagement with AI in biomedicine. The intricate and highly dynamic nature of biological systems presents a humbling challenge. Unlike well-defined mathematical problems, biological questions are multilayered, interconnected, and subject to numerous variables. In this environment, AI, with its unparalleled capacity for data processing and pattern recognition, can only accelerate our innovation engines if we, the scientists, are exceptionally adept at posing the most relevant questions. Without precisely framed inquiries, AI might generate a deluge of statistically significant correlations that lack biological meaning or clinical utility – a "42" without an understanding of the question.

AI may never be able to fully replace human scientific intuition, nor the cautious validation work that has secured the discovery of clinically successful drugs in decades past. Nevertheless, rigorously trained AI models can serve as an indispensable scientific partner and a powerful instrument that amplifies and enriches our scientific capabilities, enabling us to analyze vast datasets, make good predictions,

and thus generate hypotheses at a scale and speed impossible for humans alone. This collaborative cross talk between human intelligence inventing new paradigms (such as nodal biology) and artificial intelligence will help us find the right targets or nodes connecting many human diseases (component one), build the best drug for each of these nodal targets (component two), and ultimately give each drug to the right patients for maximum benefit (component three). Of course, no drug can be developed and moved into human clinical studies without thorough prior validation using complex systems such as human organoids (human mini-organs grown in the laboratory) or animals.¹¹ However, the speed and cost of these validation studies can be markedly reduced if we have higher confidence in the drug targets that are moved forward to validation. If accurate predictions using AI can help us increase the success rate of target validation studies and the clinical trials that follow, such that overall drug development success rates increase from 13 percent today to above 30 percent in the decade ahead, the benefits to humanity will be incalculable, getting us closer than ever to our goal of curing as many diseases as possible in our lifetimes.

I began this essay with the ambition to conceptualize a framework with which to understand the inner workings of human cells in health and disease and enable effective methods to *predict* their responses to any disease-causing perturbation. I am convinced that solving the cell perturbation prediction problem will allow us to ultimately integrate it into models of entire organs and eventually the human body. Success in these efforts will revolutionize biomedicine: the ability to predict and control the behavior of human cells will radically enhance our ability to diagnose diseases (detect disruptions in cells) and, more important, treat them in an exquisitely targeted fashion (direct drugs to specific cell disruptions). This is how we will build the drug discovery innovation engine of the future, starting today. After all, as Jules Verne is quoted as saying, “all that is impossible remains to be accomplished.”

AUTHOR'S NOTE

This essay was researched in collaboration with Gemini 3 Pro. Every word is endorsed by the author.

ABOUT THE AUTHOR

Anna Greka is a physician-scientist, a Professor at Harvard Medical School and Mass General Brigham, and a Core Institute Member of the Broad Institute of MIT and Harvard, where she also serves as Founder and Chair of the Ladders to Cures Accelerator and a Member of the Institute's Executive Leadership Team. Her work has been published in *Science*, *Cell*, and *Nature*.

ENDNOTES

- ¹ John Jumper, Richard Evans, Alexander Pritzel, et al., "Highly Accurate Protein Structure Prediction with AlphaFold," *Nature* 596 (7873) (2021): 583–589; and Josh Abramson, Jonas Adler, Jack Dunger, et al., "Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3," *Nature* 630 (8016) (2024): 493–500.
- ² Nicole M. Gaudelli, Alexis C. Komor, Holly A. Rees, et al., "Programmable Base Editing of A·T to G·C in Genomic DNA without DNA Cleavage," *Nature* 551 (7681) (2017): 464–471.
- ³ Moran Dvela-Levitt, Maria Kost-Alimova, Maheswarareddy Emani, et al., "Small Molecule Targets TMED9 and Promotes Lysosomal Degradation to Reverse Proteinopathy," *Cell* 178 (8) (2019): 521–535; and Jillian L. Shaw, Juan Lorenzo Pablo, and Anna Greka, "Mechanisms of Protein Trafficking and Quality Control in the Kidney and Beyond," *Annual Review of Physiology* 85 (2023): 407–423.
- ⁴ Petra Kaufmann, Anne R. Pariser, and Christopher Austin, "From Scientific Discovery to Treatments for Rare Diseases—The View from the National Center for Advancing Translational Sciences—Office of Rare Diseases Research," *Orphanet Journal of Rare Diseases* 13 (2018): 196; The Lancet Global Health, "The Landscape for Rare Diseases in 2024," *The Lancet Global Health* 12 (3) (2024): e341; and Global Genes, "Numbers: RARE Disease Facts," <https://globalgenes.org/rare-disease-facts> (accessed December 22, 2025).
- ⁵ Vlado Perkovic, Jonathan Barratt, Brad Rovin, et al., "Alternative Complement Pathway Inhibition with Iptacopan in IgA Nephropathy," *The New England Journal of Medicine* 392 (6) (2025): 531–543; and Kate McKeage, "Ravulizumab: First Global Approval," *Drugs* 79 (3) (2019): 347–352.
- ⁶ Dvela-Levitt, Kost-Alimova, Emani, et al., "Small Molecule Targets TMED9 and Promotes Lysosomal Degradation to Reverse Proteinopathy"; and Shaw, Lorenzo Pablo, and Greka, "Mechanisms of Protein Trafficking and Quality Control in the Kidney and Beyond."
- ⁷ Gaudelli, Komor, Rees, et al., "Programmable Base Editing of A·T to G·C in Genomic DNA without DNA Cleavage"; Atray Dixit, Oren Parnas, Biyu Li, et al., "Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens," *Cell* 167 (7) (2016): 1853–1866; and Molly Gasperini, Andrew J. Hill, José L. McFaline-Figueroa, et al., "A Genome-Wide Framework for Mapping Gene Regulation via Cellular Genetic Screens," *Cell* 176 (1–2) (2019): 377–390.
- ⁸ Yang Xu, Stephen Fleming, Matthew Tegtmeyer, et al., "Modeling Interpretable Correspondence between Cell State and Perturbation Response with CellCap," *bioRxiv* (2024), <https://doi.org/10.1101/2024.03.14.585078>; and Mehrtash Babadi, Stephen Fleming, Yang Xu, et al., "Cellarium AI Platform," Cellarium, Broad Institute of MIT and Harvard, <https://cellarium.ai/tool/cellarium-ai-platform> (accessed December 22, 2025).

- ⁹ Andrew Kirby, Andreas Gnirke, David B. Jaffe, et al., “Mutations Causing Medullary Cystic Kidney Disease Type 1 Lie in a Large VNTR in MUC1 Missed by Massively Parallel Sequencing,” *Nature Genetics* 45 (3) (2013): 299–303.
- ¹⁰ Dvela-Levitt, Kost-Alimova, Emani, et al., “Small Molecule Targets TMED9 and Promotes Lysosomal Degradation to Reverse Proteinopathy”; and Le Xiao, Xiong Pi, Alissa C. Goss, et al., “Molecular Basis of TMED9 Oligomerization and Entrapment of Misfolded Protein Cargo in the Early Secretory Pathway,” *Science Advances* 10 (38) (2024): eadp2221.
- ¹¹ Ayshwarya Subramanian, Eriene-Heidi Sidhom, Maheswarareddy Emani, et al., “Single Cell Census of Human Kidney Organoids Shows Reproducibility and Diminished Off-Target Cells after Transplantation,” *Nature Communications* 10 (1) (2019): 5462.

From Pixels to Minds: Mapping & Understanding the Brain with AI

Viren Jain & Jeff Lichtman

Enabled by advances in imaging and computer science, the past two decades have featured dramatic progress in the study of comprehensive synaptic-resolution maps of the nervous system. In this essay, we review these achievements and innovations but also argue that subsequent progress will be substantially driven by artificial intelligence. Specifically, we propose that AI is poised to revolutionize the study of brain wiring by: 1) helping to forge a new, functionally grounded definition of “understanding” the brain; 2) enabling sophisticated simulations and predictive models of neural circuits; and 3) identifying subtle connectomic “fingerprints” of neurological and psychiatric diseases. We envision a future in which the joint contributions of connectomics and AI not only decipher the brain’s subtle wiring but also unlock novel avenues for fundamental discovery and therapeutic intervention.

It is no small wonder that brains are poorly understood. In virtually all animals that have nervous systems, the brain exceeds the complexity of all other parts of the body. Because brains both inherit their structure from their ancestors and modify it by virtue of experience, a full description of how they work spans from the role of genetic information and molecular biology and biochemistry all the way to studies of psychology and ethology. If that is not enough, perhaps the deepest obstacle originates from the way brain cells – neurons – carry out their functions. Each cell is part of a neural network connected by biological wires (axons and dendrites) into a vast relay that spans behavior, from sensation to action. Neurons collect information from sensory experiences via their dendrites or from other, farther upstream neurons passing information along. Each neuron conveys what it extracts from its input to downstream neurons via its highly branched axon. The relay finally terminates with the neurons responsible for action by causing the contraction of muscle fibers that let us speak, move, write, and otherwise act.

This network is, even in the simplest organisms, much more complicated than one might have imagined. Neurons are wired not just to a few other cells but, in many cases, thousands of upstream and downstream partners. The partners are

often far flung and connected by biological wires that extend in many directions over very long distances. The sites of communication are specialized structures called synapses. The human brain has trillions of synapses. Mapping the complete synaptic connectional map, called a *connectome*, is the foundational blueprint of brain function – and this is the quest we and other researchers are on.

Until the late nineteenth century, the study of the brain was constrained by the impossibility of isolating and identifying individual cells in stains of nervous tissue. Then, in 1873, Italian biologist and pathologist Camillo Golgi discovered what would become known as the Golgi stain, a method of staining individual neurons at random, revealing the axon and dendrites of complete nerve cells and thus the structures and connections of the nervous system. Spanish anatomist Santiago Ramón y Cajal improved on this technique and contributed detailed drawings of the branched structure of the brain. Together, Golgi and Cajal showed that neurons are connected to each other in directional pathways, an insight for which they won the Nobel Prize in Physiology or Medicine.¹

But Cajal's quasidiagrammatic renderings belied a reality that is only now being revealed; his drawings showed only the purported connections that were consistent with the pathway he was trying to illustrate. In contrast, connectomics specifically intends to show *all* the connections. In this sense, connectomics data are not biased, which is a blessing and a curse. On the positive side, the dense labeling of all neurons provides a highly accurate rendering of all the inputs a neuron is receiving and reveals where that information is going via all the downstream neurons its axon establishes synapses with. On the negative side, it is no longer so straightforward to extract a specific functional pathway, given the profound complexity of the overall network.

The journey to map neural circuits has been punctuated by significant technological leaps and landmark discoveries, transforming connectomics from a theoretical aspiration into a data-rich empirical science. In biological fields such as connectomics, genomics, and transcriptomics, as well as physical sciences like astronomy, large amounts of raw data are often required, making data acquisition both a technological challenge and an opportunity for innovative solutions. In the relatively young field of connectomics, data requirements have led to many different approaches to imaging the biological structures underlying the wiring diagram of nerve cells. Because of the miniscule size of the nerve cell wires and their synapses, traditional light microscopy does not have sufficient resolving power. Attempts to overcome this limitation include a technicolor Golgi-like neuron labeling approach called "Brainbow," in which neurons are labeled with many different colors (more than fifty), as well as more recent methods involving physical enlargement of biological tissue.² However the most successful method thus far is to avoid light microscopy altogether in favor of using electrons, which have wavelengths, and hence resolutions that are at least ten thousand-fold more detailed than the photons used in light microscopy.

Two major types of electron microscopy (EM) are used in connectomics.³ In transmission electron microscopy (TEM), the imaged signal is based on detecting electrons that are transmitted through a sample. Scanning electron microscopy (SEM) approaches, on the other hand, detect electrons that are scattered from the sample's surface. These two approaches also allow for different ways of preparing tissue. It is generally accepted that tracing wires unambiguously requires distinguishing them both from nearby wires in the same plane of focus as well as from wires that lie slightly above or below them. Therefore, the easiest way to avoid blur from tissue above or below a particular neuronal wire is to physically cut brain samples into very thin slices. Automated ways of collecting ultra-thin (one-thousandth the thickness of a hair) sections onto flexible tape are used in both SEM and TEM approaches.⁴ This “section then image” approach is something akin to a film strip, with the images of physical brain sections like the frames of a movie. An alternative approach is to “image then section” a thick brain sample (often called a “block”). This technique, known as block-face imaging, uses scanning electron microscopy.⁵ The electrons come from the interaction of the scanning electron beam with the top layer of the block. Once an image is generated, the top layer is sectioned away and the block-face is scanned again; this process repeats until a volume of hundreds or thousands of consecutive block-faces have been imaged.

Other novel approaches include trying to use light (photons) rather than electron imaging. One potentially promising avenue is to use short-wavelength light (X-rays) to induce electron backscatter with rapid imaging of osmium stained brains.⁶ Visible longer-wavelength light can also be used, despite its lower resolving power, if the specimen is stretched into a larger size – a technique called “expansion microscopy.”⁷ This approach is especially exciting because it offers a straightforward way of superimposing molecular “labels” that, for example, clearly differentiate cell types in a connectome dataset.⁸

With these new tools in hand, some researchers have begun to transition from developing techniques to obtaining large-volume datasets for neurobiological research. But samples large enough to contain neuronal circuits require automated acquisition techniques. Indeed, using manual techniques, the first connectome of *C. elegans*, the millimeter-long round worm with only about three hundred neurons, took a decade.⁹ About forty years later, in 2021, eight worm connectomes were completed in two years, a forty-fold speedup.¹⁰ Beyond *C. elegans*, an intense effort is now underway to create a full description of the nervous system of *Drosophila melanogaster*, the fruit fly.¹¹ The small size of this animal and its long history as a model organism for genetic studies make it an ideal candidate to merge genetic and structural (circuit) information for a deeper understanding of the underpinnings of behavior. The late biologist and geneticist Seymour Benzer was the first to focus attention on finding the genetic basis of

behavior and used the fly as his model organism.¹² At the time, there was little if any circuit information to go on, so his focus was on genes that modify behavior. Today, now that the connectome – the physical substrate of behavior – is becoming accessible, researchers hope this field may see a resurgence. Fly connectomics has used both serial block-face and transmission electron microscopy approaches, carried out in both adults and larval instars.¹³ The brains of many other species (for example, the octopus and microwasp) are also being scrutinized.¹⁴ A vertebrate, the larval zebrafish (*Danio rerio*), is a relatively transparent, rapidly developing freshwater fish that is easily bred in captivity and amenable to large-scale genetic screens. It has also been possible to study its behavior and the activity of the neurons related to these behaviors with fluorescence microscopy of cell activity.¹⁵ Mammalian connectomics is moving swiftly. A number of groups are building connectomes of parts of the mouse brain and peripheral nervous system.¹⁶ Many labs are working to devise an approach to obtain the full connectome of a mouse brain. This is a daunting task, as a cubic millimeter is far less than 1 percent of a whole mouse brain, vaulting connectomics from the petascale to the exascale.¹⁷ Meeting this challenge will require both far more automation and ramping up computational and AI tools.

Finally, human connectomics has recently got off the ground with the acquisition and analysis of petascale cortex datasets, revealing novel circuit organization properties (such as much stronger pairwise neuron connections than expected) and some truly unexpected anatomical structures (like axons that tie themselves into knots).¹⁸ Despite the significant imaging and computational achievements associated with such data, these cubic millimeter-scale data represent only one-millionth of an entire human brain. Whatever the future of human connectomics holds, the endeavor is clearly just beginning.

The sheer volume and complexity of connectomic image data render manual analysis an insurmountable bottleneck. AI, particularly neural network-based deep-learning methods, has been crucial in transforming raw imaging data (pixels) into meaningful, reconstructed neural pathways.

The most labor-intensive step in early connectomics was the manual tracing of neuronal processes (segmentation) and the identification of synapses across thousands of electron microscopy images. AI has revolutionized this domain. Early computational methods provided minimal assistance, but the advent of deep learning, particularly convolutional neural networks, convolutional outgrowths like U-Nets, and more specialized architectures like flood-filling networks, has enabled highly accurate automated neurite segmentation and synapse detection.¹⁹ These algorithms can learn the complex visual patterns associated with cell membranes and organelles like mitochondria and synaptic vesicles directly from human-annotated training data, achieving performance levels that, while not yet perfect,

significantly reduce the need for manual intervention, thus speeding the segmentation process profoundly.

Beyond reconstruction, AI is also proving vital for the initial analysis of connectomic data. For instance, machine-learning algorithms, including self-supervised learning methods, are being used for the automated classification of cell types based on their intricate morphology and detailed connectivity patterns.²⁰ This allows for a more objective and scalable approach to cataloging the brain's cellular constituents, moving beyond traditional, often subjective, classification schemes.

This work has produced a number of insights that challenge our existing views of the ways brains work. For example, the diversity of neuronal structures and connectivity in individual animals is astonishing; in *Drosophila*, there may be more than a thousand different types of neurons, each of which has its own structure, gene expression profile, and connectivity rules.

Therefore, while mapping the structural connectome – the graph of synaptic connections – is a monumental and critical step forward, it is becoming clear that a purely structural map, however detailed, provides an incomplete picture of brain function. Neurons are not just nodes in a graph; they are dynamic, molecularly diverse entities.

A critical limitation of traditional connectomics is the absence of information about the molecular and physiological diversity inherent in neural circuits. The brain utilizes a vast set of neurotransmitters, each acting on a similarly diverse array of receptors, leading to varied effects on neural activity. Ion channels, with their distinct biophysical properties, shape neuronal excitability and signaling. Furthermore, neuromodulatory systems that release substances like dopamine, serotonin, or acetylcholine can globally reconfigure circuit states and synaptic efficacy, often without overt structural changes detectable by standard microscopy.²¹ These molecular dimensions are largely invisible in purely anatomical reconstructions.

To address these limitations, the field is moving toward “multimodal” connectomics. This involves integrating data from various “omics” disciplines – genomics, transcriptomics, and proteomics – with structural connectomic maps to create a “molecular connectome.”²² Pioneering imaging approaches are being developed to generate such multimodal data. Correlative light and electron microscopy (CLEM) allows researchers to link functional or molecular information obtained with light microscopy to ultrastructural details from EM.²³ Techniques like light microscopy – based connectomics (LICONN) use expansion microscopy and light microscopy to visualize synaptic connectivity and molecular components in the same tissue.²⁴ AI is set to play a crucial role in fusing these disparate data types. Early successes are emerging in using AI for cross-modal understanding and prediction, such as inferring molecular features or cell types directly from EM image data, leveraging the subtle textural and morphological cues that AI models can learn.²⁵ This AI-enabled integration will vastly improve the interpretability of connectomic data.

The accumulation of vast connectomic datasets, while a triumph of modern neuroscience and engineering, presents a new, even more formidable challenge: interpretation. How do we move from petabytes of anatomical data to a genuine understanding of how neural circuits compute, learn, and generate behavior?

Consider the history of *C. elegans* research. Although it might seem straightforward to relate neural circuitry to behavior in an organism that has only three hundred brain cells, this ambition has been much more challenging than expected. Indeed, the original *C. elegans* connectome project was part of an ambitious plan to relate genetics, the nervous system, and behavior. This effort led to several landmark discoveries concerning the lineage of the cells that make up the nervous system, the molecules that underlie the naturally occurring death of cells, and, of course, the first whole organism connectome. But a grand synthesis of the crucial strands of patterns of gene expression, neural development, neural wiring, and behavior has been elusive. The problem is complexity. All the aspects of an animal's physiology are honed by millions of years of evolution. Over time, "accidents" due to random mutations become ingenious tweaks, or sometimes less elegant kludges, embedded in the genetic makeup of an animal. Biological systems are thus relatively difficult to reverse engineer from the point of view of a simple, human-interpretable model or explanation derived by manual analysis. We therefore argue that AI is essential not only for managing and reconstructing data but for unlocking its meaning.

The very notion of "understanding" a system as complex as the brain is multifaceted and has evolved with our scientific capabilities. Historically, understanding might have meant a detailed anatomical description or a qualitative model of information flow. However, in the context of connectomics and AI, we propose a more rigorous and functional definition of understanding.

A crucial aspect of this redefinition involves moving beyond purely descriptive cataloging of connections and cell types. While such maps are foundational, a more powerful form of understanding implies the ability to make predictions about the system's behavior and ultimately to generate or simulate aspects of its function. This shifts the benchmark for understanding toward the development of causal, predictive, and generative models of neural circuits.

Under this framework, "understanding" a neural circuit could be benchmarked by the ability to:

1. *Accurately simulate circuit activity.* Given a structural connectome and relevant physiological parameters (some of which might be inferred by AI), can we simulate the patterns of neural activity that would arise under diverse physiological inputs or pathological conditions?²⁶
2. *Predict the functional consequences of perturbations.* Can we predict how the circuit's output or internal dynamics would change if specific neurons

were silenced, connections were severed (as in a lesion study), or synaptic strengths were altered?

3. *Generate realistic synthetic connectomes.* Can AI models, trained on existing connectomic data, learn the underlying “design principles” or “grammatical rules” of neural wiring to the point at which they can generate novel, synthetic connectomes that are statistically and perhaps even functionally indistinguishable from biological ones?

One of the most exciting prospects of integrating AI with connectomics is the ability to create large-scale, biologically constrained simulations of neural circuits. These *in silico* experiments can complement and guide *in vivo* and *in vitro* studies, allowing for rapid hypothesis testing and exploration of parameter spaces that would be intractable experimentally.

AI models can be trained to forecast network activity, information flow, and emergent properties based on structural connectomes. For instance, given the wiring diagram of a cortical column, an AI model might predict how patterns of sensory input propagate through the layers, how different cell types contribute to a computation, or how network oscillations arise. Benchmarks for evaluating such predictive models will be crucial for driving progress and comparing different modeling approaches.²⁷ These models can be employed for *in silico* hypothesis testing regarding circuit mechanisms, such as by simulating the effect of blocking a particular receptor type or altering the firing properties of a specific interneuron class. A significant challenge is to develop models that can robustly predict functional connectivity (patterns of correlated activity) from detailed structural connectomes, bridging the gap between anatomical maps and dynamic brain states.

Connectomic data provide an unprecedented causal foundation and set of modeling constraints for comprehensive neural simulations.²⁸ However, structure alone is insufficient; physiological parameters such as synaptic weights, channel conductances, and cell-intrinsic properties are also needed. AI methodologies offer a path to inferring these missing parameters. For example, machine learning models could be trained on datasets combining sparse functional recordings (for instance, multi-electrode arrays or calcium imaging) with dense connectomic reconstructions from the same tissue volume. Computational models could learn to estimate synaptic strengths or neuronal excitability profiles that best reproduce the observed functional data, given the anatomical constraints. This AI-driven fusion of multi-modal data, integrating electrophysiology, activity imaging, and even molecular profiles (such as from spatial transcriptomics correlated with connectomic data), is a key area for future development.

The brain’s wiring is shaped by evolution, development, and experience to perform sophisticated computations. AI can help discover hidden principles, motifs, and rules that govern this “neural syntax.”

The connectome is fundamentally a graph, with neurons as nodes and synapses as edges. Traditional graph theory has provided initial insights, but the sheer scale and complexity of mammalian connectomes demand more powerful analytical tools. Graph neural networks and other geometric deep-learning approaches are well suited for analyzing such complex relational data. These AI models can learn to identify higher-order connectivity motifs (patterns involving multiple neurons and synapses), circuit invariants (recurring structural features across different brain regions or even species), and fundamental organizational principles that might not be apparent through manual inspection or simpler statistical analyses. Furthermore, unsupervised and self-supervised learning techniques can be applied to discover novel structural and functional features within connectomic data without requiring *a priori* hypotheses or explicit labels, potentially revealing entirely new ways of categorizing neurons or understanding circuit organization.

Evolution has sculpted nervous systems to meet diverse behavioral demands. Comparing connectomes across different species, developmental stages, learning paradigms, or even between sexes can reveal conserved computational solutions and species-specific adaptations. AI can be instrumental in this comparative analysis. For instance, researchers may develop and employ AI tools for the robust alignment and comparison of massive, complex connectomic datasets, even when there isn't a one-to-one correspondence between neurons. This can help quantify interindividual variability in connectomes and its functional implications, a key research area for understanding both normal brain function and the basis of individual differences.²⁹ Identifying conserved circuit motifs that appear across different phyla might point to fundamental computational building blocks, while divergent features could explain unique cognitive abilities.

Many neurological and psychiatric disorders are increasingly thought to have underpinnings in altered neural circuitry, a concept often termed “connectopathies.”³⁰ AI-supported synapse-level connectomics offers the potential to identify subtle structural “fingerprints” of these conditions, paving the way for new diagnostic tools and a deeper understanding of disease mechanisms.

AI algorithms can be trained to detect subtle, disease-associated alterations in synaptic density, the morphology of dendritic spines or axons, or specific patterns of miswiring in animal models of disease or, where available, in human pre- or postmortem tissue.³¹ For example, an AI system could learn to distinguish the synaptic architecture in a mouse model of autism from that of wild-type control (unmodified) mice, even if the differences are too subtle or distributed for a human pathologist to reliably identify. This enables high-throughput screening of connectomic changes, potentially leading to the discovery of novel biomarkers for early diagnosis, disease stratification, or monitoring treatment response.

Beyond biomarker discovery, AI-driven simulations based on connectomes from disease models can help investigate how specific connectomic aberrations lead to

functional deficits. For instance, if a particular interneuron type is found to have reduced synaptic input in a schizophrenia model, simulations could explore how this local change impacts broader network dynamics and information processing, potentially linking the connectomic alteration to cognitive symptoms. Such models could also be used to develop predictive tools for disease progression or to simulate the effects of potential therapeutic interventions that aim to correct the underlying circuit pathology.

A significant hurdle in applying connectomics to human disease is the limited availability of synapse-level human brain tissue suitable for EM, especially from living patients or for psychiatric conditions with which postmortem changes can be confounding. The recent Ho1 dataset from a human cortical sample – a 1.4 petabyte reconstruction of one cubic millimeter of human brain tissue – while from a patient with epilepsy, represents a step forward, but continued efforts to acquire and analyze such samples, combined with AI's ability to extract maximal information from them, will be crucial.³²

The brain is the only existing proof that general, robust, and energy-efficient intelligence is possible. As connectomics unveils the detailed circuit diagrams of biological brains, it will provide a rich source of inspiration for new AI architectures and algorithms. Insights into how neurons are wired, the types of circuit motifs employed, principles of sparsity and wiring economy, and the structural basis of learning rules could directly inform the design of next-generation AI systems that are more powerful, energy-efficient, data-efficient, and biologically plausible.³³ For instance, understanding the specific connectivity patterns in brain regions known for particular cognitive functions (like navigation, language, or planning) might lead to AI models that excel at these tasks by mimicking biological solutions.

The synergy also flows in the reverse direction. As AI research develops increasingly powerful architectures, such as transformers, the fascinating question arises: Do biological neural circuits implement analogous computational principles?³⁴ Large-scale connectomics provides the data to investigate this. By analyzing detailed wiring diagrams, neuroscientists can search for structural motifs or patterns of connectivity that might correspond to components of successful AI models. For example, are there patterns of long-range connections and local inhibitory circuits in the cortex that could implement something akin to the artificially engineered attention mechanisms in transformers? Discovering such homologies would not only deepen our understanding of brain function but could also help refine AI concepts.

The prospect of acquiring and analyzing highly detailed brain data, potentially including synapse-level connectomes from humans, raises significant ethical considerations that must be proactively addressed.

Human connectomic data are arguably among the most personal and sensitive information imaginable, especially if and when such data are generated at significant scale. Robust frameworks for data privacy, secure storage, and controlled access are essential. Questions of data ownership – whether by the individual, the research institution, or the funding agency – need careful consideration, especially as datasets become more widely shared for collaborative research.

As AI tools identify potential connectomic biomarkers for complex traits or disorders, there is also a risk of misinterpretation or oversimplification. Brain function and behavior are highly complex, arising from interactions across multiple scales. Attributing a complex psychiatric condition solely to a specific wiring abnormality identified by AI would be a gross oversimplification and could lead to stigmatization or flawed therapeutic approaches. Clear communication of the limitations and probabilistic nature of such findings is crucial.

We must also consider the profound philosophical and societal implications of the ability to deeply understand, simulate, and potentially, in the distant future, modify brain circuits. What does it mean for our concepts of self, free will, and consciousness if we can create detailed computational replicas of brain activity? While therapeutic applications for devastating brain disorders are a primary motivator, the potential for nontherapeutic uses or misuse of such powerful technologies cannot be ignored. Open discussion involving scientists, ethicists, policymakers, and the public will be necessary to navigate these complex issues responsibly.

Connectomics has advanced from its highly conceptual origins to its current state as a data-rich, technologically advanced field. The initial, painstaking efforts to map the nervous system of *C. elegans* laid a crucial foundation, demonstrating the feasibility and potential value of complete wiring diagrams. Over the past two decades, driven by exponential improvements in electron microscopy, automated sectioning, and computational power, the scale and scope of connectomic projects have expanded dramatically, culminating in recent landmark datasets from *Drosophila*, mice, and even fragments of the human brain. Throughout this evolution, machine learning and AI have transitioned from a peripheral aid to a central and indispensable partner. Initially applied to the formidable challenge of segmenting neurons and synapses from vast image volumes, AI's role has broadened to encompass more detailed forms of data annotation.

Now, however, AI is poised to play an even more fundamental role in scientific interpretation of brain data. Specifically, AI offers the tools to redefine what it means to “understand” the brain, moving toward predictive and generative models that can be rigorously tested and iteratively refined. It allows us to perform *in silico* experiments at scales and with precision previously unimaginable, to uncover hidden organizational principles within the enormous complexity of neural wiring,

and to identify the subtle connectomic fingerprints that may unlock the mysteries of devastating brain disorders.

Looking forward, the potential for paradigm-shifting breakthroughs is immense. The coevolution of connectomics and AI promises not only to accelerate discoveries in fundamental neuroscience but also to inspire new generations of artificial intelligence, potentially leading to AI systems that are more robust and efficient. In the realm of medicine, the ability to link specific connectomic alterations to disease states could substantially accelerate diagnostics and pave the way for novel therapeutic strategies targeted at the circuit level.

However, realizing this vision requires a sustained and intensified commitment to interdisciplinary collaboration. The challenges ahead are too vast and complex for any single discipline or entity to tackle alone. Neuroscientists, AI researchers, computer scientists, engineers, physicists, mathematicians, and ethicists must work in concert, sharing data, tools, and insights. Initiatives that foster such “big team science,” alongside continued investment in foundational technologies and computational infrastructure, will be critical.

ABOUT THE AUTHORS

Viren Jain is Senior Staff Research Scientist at Google, where he leads the Connectomics team. He also serves on the Scientific Advisory Council of the Allen Institute for Brain Science and on the Advisory Board for the Chan-Zuckerberg Initiative Imaging program. He has recently published in such journals as *Nature*, *Cell*, and *Science*.

Jeff Lichtman is the Jeremy R. Knowles Professor of Molecular and Cellular Biology and the Santiago Ramón y Cajal Professor of Arts and Sciences at Harvard University. He has recently published in such journals as *Nature Methods*, *Current Biology*, and *IEEE Journal of Biomedical and Health Informatics*.

ENDNOTES

¹ Santiago Ramón y Cajal, “The Structure and Connexions of Neurons: Nobel Lecture, December 12, 1906,” The Nobel Prize, <https://www.nobelprize.org/uploads/2018/06/cajal-lecture.pdf>.

² Fred Y. Shen, Margaret M. Harrington, Logan A. Walker, et al., “Light Microscopy Based Approach for Mapping Connectivity with Molecular Specificity,” *Nature Communications* 11 (1) (2020); and Mojtaba R. Tavakoli, Julia Lyudchik, Michał Januszewski, et al., “Light-Microscopy-Based Connectomic Reconstruction of Mammalian Brain Tissue,” *Nature* 642 (8067) (2025): 398–410.

- ³ Kevin L. Briggman and David D. Bock, “Volume Electron Microscopy for Neuronal Circuit Reconstruction,” *Current Opinion in Neurobiology* 22 (1) (2012): 154–161.
- ⁴ Narayanan Kasthuri, Kenneth Jeffrey Hayworth, Daniel Raimund Berger, et al., “Saturated Reconstruction of a Volume of Neocortex,” *Cell* 162 (3) (2015): 648–661; and Jasper S. Phelps, David Grant Colburn Hildebrand, Brett J. Graham, et al., “Reconstruction of Motor Control Circuits in Adult *Drosophila* Using Automated Transmission Electron Microscopy,” *Cell* 184 (3) (2021): 759–774.
- ⁵ Winfried Denk and Heinz Horstmann, “Serial Block-Face Scanning Electron Microscopy to Reconstruct Three-Dimensional Tissue Nanostructure,” *PLOS Biology* 2 (11) (2004), <https://doi.org/10.1371/journal.pbio.0020329>.
- ⁶ Kevin M. Boergens, Gregg Wildenberg, Ruiyu Li, et al., “Photoemission Electron Microscopy for Connectomics,” *bioRxiv* (2024), <https://doi.org/10.1101/2023.09.05.556423>.
- ⁷ Asmamaw T. Wassie, Yongxin Zhao, and Edward S. Boyden, “Expansion Microscopy: Principles and Uses in Biological Research,” *Nature Methods* 16 (1) (2019): 33–41.
- ⁸ Tavakoli, Lyudchik, Januszewski, et al., “Light-Microscopy-Based Connectomic Reconstruction of Mammalian Brain Tissue.”
- ⁹ John Graham White, Eileen Southgate, J. N. Thomson, and Sydney Brenner, “The Structure of the Nervous System of the Nematode *Caenorhabditis elegans*,” *Philosophical Transactions of the Royal Society B: Biological Sciences* 314 (1165) (1986): 1–340.
- ¹⁰ Daniel Witvliet, Ben Mulcahy, James K. Mitchell, et al., “Connectomes across Development Reveal Principles of Brain Maturation,” *Nature* 596 (7871) (2021): 257–261.
- ¹¹ Louis K. Scheffer, C. Shan Xu, Michal Januszewski, et al., “A Connectome and Analysis of the Adult Central Brain,” *Elife* 9 (2020), <https://doi.org/10.7554/elife.57443>; and Sven Dorkenwald, Arie Matsliah, Amy R. Sterling, et al., “Neuronal Wiring Diagram of an Adult Brain,” *Nature* 634 (8032) (2024): 124–138.
- ¹² Seymour Benzer, “From the Gene to Behavior,” *JAMA* 218 (7) (1971): 1015–1022.
- ¹³ Zhihao Zheng, J. Scott Lauritzen, Eric Perlman, et al., “A Complete Electron Microscopy Volume of the Brain of Adult *Drosophila melanogaster*,” *Cell* 174 (3) (2018): 730–743, <https://doi.org/10.1016/j.cell.2018.06.019>; Zhiyuan Lu, C. Shan Xu, Kenneth J. Hayworth, et al., “En bloc Preparation of *Drosophila* Brains Enables High-Throughput FIB-SEM Connectomics,” *Frontiers in Neural Circuits* 16 (2022): 917251; and Michael Winding, Benjamin D. Pedigo, Christopher L. Barnes, et al., “The Connectome of an Insect Brain,” *Science* 379 (6636) (2023), <https://doi.org/10.1126/science.add9330>.
- ¹⁴ Flavie Bidel, Yaron Meirovitch, Richard Lee Schalek, et al., “Connectomics of the *Octopus vulgaris* Vertical Lobe Provides Insight into Conserved and Novel Principles of a Memory Acquisition Network,” *eLife* 12 (2023), <https://doi.org/10.7554/eLife.84257>; and Nicholas J. Chua, Anastasia A. Makarova, Pat Gunn, et al., “A Complete Reconstruction of the Early Visual System of an Adult Insect,” *Current Biology* 33 (21) (2023): 4611–4623.
- ¹⁵ Misha B. Ahrens, Michael B. Orger, Drew N. Robson, et al., “Whole-Brain Functional Imaging at Cellular Resolution Using Light-Sheet Microscopy,” *Nature Methods* 10 (5) (2013): 413–420; and Jan-Matthis Lueckmann, Alexander Immer, Alex Bo-Yuan Chen, et al., “ZAPBench: A Benchmark for Whole-Brain Activity Prediction in Zebrafish,” in *The 13th International Conference on Learning Representations (ICLR, 2024)*.

- ¹⁶ Alessandro Motta, Manuel Berning, Kevin M. Boergens, et al., “Dense Connectomic Reconstruction in Layer 4 of the Somatosensory Cortex,” *Science* 366 (6469) (2019), <https://doi.org/10.1126/science.aay3134>; Davi D. Bock, Wei-Chung Allen Lee, Aaron M. Kerlin, et al., “Network Anatomy and *in vivo* Physiology of Visual Cortical Neurons,” *Nature* 471 (2011): 177–182; and Ju Lu, Juan Carlos Tapia, Olivia L. White, and Jeff W. Lichtman, “The Interscutularis Muscle Connectome,” *PLOS Biology* 7 (2) (2009), <https://doi.org/10.1371/journal.pbio.1000032>.
- ¹⁷ Larry F. Abbott, Davi D. Bock, Edward M. Callaway, et al., “The Mind of a Mouse,” *Cell* 182 (6) (2020): 1372–1376.
- ¹⁸ Alexander Shapson-Coe, Michał Januszewski, Daniel R. Berger, et al., “A Petavoxel Fragment of Human Cerebral Cortex Reconstructed at Nanoscale Resolution,” *Science* 384 (6696) (2024), <https://doi.org/10.1126/science.adk4858>.
- ¹⁹ Yann LeCun, Bernhard Boser, John S. Denker, et al., “Backpropagation Applied to Handwritten Zip Code Recognition,” *Neural Computation* 1 (4) (1989): 541–551; Viren Jain, Joseph F. Murray, Fabian Roth, et al., “Supervised Learning of Image Restoration with Convolutional Networks,” in 2007 *IEEE 11th International Conference on Computer Vision* (IEEE, 2007), <https://ieeexplore.ieee.org/document/4408909>; Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, ed. Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (Springer, 2015), 234–241; Kisuk Lee, Jonathan Zung, Peter Li, et al., “Superhuman Accuracy on the SNEMI3D Connectomics Challenge,” arXiv (2017), <https://doi.org/10.48550/arXiv.1706.00120>; and Michał Januszewski, Jörgen Kornfeld, Peter H. Li, et al., “High-Precision Automated Reconstruction of Neurons with Flood-Filling Networks,” *Nature Methods* 15 (8) (2018): 605–610.
- ²⁰ Sven Dorkenwald, Peter H. Li, Michał Januszewski, et al., “Multi-Layered Maps of Neuropil with Segmentation-Guided Contrastive Learning,” *Nature Methods* 20 (12) (2023): 2011–2020.
- ²¹ Cornelia I. Bargmann and Eve Marder, “From the Connectome to Brain Function,” *Nature Methods* 10 (6) (2013): 483–490.
- ²² Jeff W. Lichtman and Joshua R. Sanes, “Ome Sweet Ome: What Can the Genome Tell Us about the Connectome?” *Current Opinion in Neurobiology* 18 (3) (2008): 346–353.
- ²³ Tao Fang, Xiaotang Lu, Daniel Berger, et al., “Nanobody Immunostaining for Correlated Light and Electron Microscopy with Preservation of Ultrastructure,” *Nature Methods* 15 (12) (2018): 1029–1032; and Xiaomeng Han, Xiaotang Lu, Peter H. Li, et al., “Multiplexed Volumetric CLEM Enabled by scFvs Provides Insights into the Cytology of Cerebellar Cortex,” *Nature Communications* 15 (1) (2024): 6648.
- ²⁴ Tavakoli, Lyudchik, Januszewski, et al., “Light-Microscopy-Based Connectomic Reconstruction of Mammalian Brain Tissue.”
- ²⁵ Dorkenwald, Li, Januszewski, et al., “Multi-Layered Maps of Neuropil with Segmentation-Guided Contrastive Learning”; and Nils Eckstein, Alexander Shakeel Bates, Andrew Champion, et al., “Neurotransmitter Classification from Electron Microscopy Images at Synaptic Sites in *Drosophila melanogaster*,” *Cell* 187 (10) (2024): 2574–2594, <https://doi.org/10.1016/j.cell.2024.03.016>.
- ²⁶ Lueckmann, Immer, Chen, et al., “ZAPBench: A Benchmark for Whole-Brain Activity Prediction in Zebrafish”; and Viren Jain, “How AI Could Lead to a Better Understanding

- of the Brain,” *Nature* 623 (7986) (2023): 247–250, <http://dx.doi.org/10.1038/d41586-023-03426-3>.
- ²⁷ Lueckmann, Immer, Chen, et al., “ZAPBench: A Benchmark for Whole-Brain Activity Prediction in Zebrafish.”
- ²⁸ Lappalainen Janne, Tschopp Fabian, Prakhya Sridhama, et al., “Connectome-Constrained Networks Predict Neural Activity Across the Fly Visual System,” *Nature* 634 (2024): 1132–1140.
- ²⁹ Joshua L. Morgan and Jeff W. Lichtman, “Why Not Connectomics?” *Nature Methods* 10 (6) (2013): 494–500.
- ³⁰ Mikail Rubinov and Ed Bullmore, “Fledgling Pathoconnectomics of Psychiatric Disorders,” *Trends in Cognitive Sciences* 17 (12) (2013): 641–647.
- ³¹ Shapson-Coe, Januszewski, Berger, et al., “A Petavoxel Fragment of Human Cerebral Cortex Reconstructed at Nanoscale Resolution.”
- ³² Ibid.
- ³³ Abbott, Bock, Callaway, et al., “The Mind of a Mouse.”
- ³⁴ Ashish Vaswani, Noam Shazeer, Niki Parmar, et al., “Attention Is All You Need,” in *NIPS ’17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, ed. Ulrike von Luxburg, Isabelle Guyon, Samy Bengio, et al. (Curran Associates, Inc., 2017).

The Algorithmic Planet

Anna M. Michalak & John C. Platt

To ensure a sustainable future, we need to understand how Earth's climate has changed over time, how different factors have contributed to those changes, and how human action will impact the climate in the future. Developing this understanding involves a continual process of model refinement, calibration, validation, and evaluation against available observations. These tasks present the core opportunity and challenge of applying artificial intelligence to climate: AI is enabling a revolution in the ability to represent the functioning of complex systems, but while the sheer volume of data available to help us understand the earth system is growing at an unprecedented pace, this observational record is often ill-suited to provide robust benchmarks for AI-driven models. In this essay, we present examples that illustrate this tension, focusing on the AI tractability of different applications, with tractability linked to the availability of metrics and benchmarks to guide model development. We also describe a future when developments in AI will accelerate improvement in models used to support climate action and resilience, enabling us to tackle the currently "intractable" frontier.

We, the authors of this essay, have up to this point spent our careers in different worlds – namely, academia and technology – but we share two things in common. The first is that we have both been working to understand and address the climate challenge for many years. The second is that we are both strong believers in developing and using tools that are firmly grounded in data. In fact, for both of us, some of our earliest papers dealt with constrained optimization, which is the process of finding the best solution of a function, called the objective function, while adhering to certain limitations or constraints.¹

The idea of constrained optimization, very broadly defined, is central to understanding Earth's past and future climate (climate science), to identifying and scaling solutions that limit the climate impact of human activity (climate mitigation), and to identifying and scaling solutions that reduce the negative impacts of climate change on societies and ecosystems (climate adaptation). In each of these cases, the goal is to maximize how much we understand about how the earth system functions under specific conditions, and act on that understanding, while dealing with the uncertainty that stems from the fact that we have limited observations on which to base that understanding.

Central to climate science, for example, is the need to understand how Earth's climate has changed over time, the degree to which different factors have contributed to those changes, and how climate will continue to change in the future as a result of what we do as humans.

Addressing even the simplest questions within the broad scope of climate science requires us to build models of Earth's functioning. For example, how do we know that Earth's temperature has risen by almost 1.5 degrees Celsius since the preindustrial era?² We can't take the temperature of Earth, as a physician would of a patient. Instead, we rely on a combination of instrumental records (like thermometers and weather satellites) and proxy data (such as ice cores and tree rings) that are unevenly distributed in space and time. Our best estimate of the global temperature is based on interpolating and extrapolating instrumental records and using models of how the earth system functions to convert proxy data into temperature estimates.

Developing models that can reconstruct historical climate conditions and anticipate future conditions, such as those used as part of the Intergovernmental Panel on Climate Change (IPCC) process, involves a continual process of model refinement and model calibration, validation, and evaluation against available observations.³ The data and observations that we have to inform projections and reconstructions are limited, however, which makes model evaluation against a comprehensive set of metrics and benchmarks challenging. As a result, even basic aspects of climate science, such as quantifying the effect of a doubling of carbon dioxide concentrations on global average temperature (known as the climate sensitivity), are still debated among climate scientists.⁴ This is because, unlike in a laboratory setting, we cannot run an experiment in which we increase the concentration of atmospheric carbon dioxide on the whole planet while keeping all other factors constant (we only have one Earth!), and our understanding of how different factors have changed concurrently over time (such as carbon dioxide concentrations, temperature, cloud cover, land cover, and ocean circulation) is limited by the lack of observations. Identifying and scaling solutions for climate mitigation and climate adaptation is similarly predicated on leveraging limited observations of the real world to constrain models of how the earth system responds to human action.

This sets up both the core opportunity and the core challenge of applying AI to climate science, climate mitigation, and climate adaptation. As described throughout this issue of *Dædalus*, artificial intelligence is enabling a revolution in the ability to represent the functioning of complex systems and in deriving understanding from massive troves of data. In many fields – such as in weather forecasting, where AI models are outperforming traditional dynamical models, and in medical imaging, where AI models are enhancing diagnostic accuracy, efficiency,

and accessibility – AI is already revolutionizing science and its applications.⁵ This presents a growing opportunity for climate science, too. The challenge, however, is that the fields in which AI has made the greatest mark are those that provide massive volumes of data on which to train models and clear benchmarks for evaluating model performance. In the case of weather, data from constellations of satellites and *in situ* instruments can be used both to train models and to assess whether forecasts were accurate. In the case of medical imaging, diagnoses can be evaluated against patient outcomes.

How does this apply to climate? While the sheer volume of data available to help us understand the earth system and our role in it is growing at an unprecedented pace, this rich observational record is often insufficient to provide robust benchmarks for AI-driven models. For example, the limited duration of data records is incompatible with the multidecadal timescales represented by climate models, the spatial and temporal scales of observations are often incompatible with models used to support climate mitigation and adaptation, the parameters that we are most interested in are often the most difficult to monitor at scale, and observational uncertainties are often too high to support model evaluation.

In what follows, we use several case studies to illustrate the opportunities and challenges associated with using artificial intelligence to inform climate action. These examples fall on different points along two axes: AI tractability and the spectrum of climate adaptation to mitigation.

The success of artificial intelligence in advancing solutions depends substantially on the tractability of specific problems and applications. AI tractability ranges from highly tractable applications, characterized by well-defined problems and supported by good data and clear benchmarks, to problems with low tractability, where the problems are harder to define, where data are sparse, where there are deep uncertainties, or where systems are nonstationary. In effect, AI tractability expresses how well suited a problem is to current AI methods. We find that tractability is linked to the availability of metrics and benchmarks to guide model development.⁶

The second axis represents the climate adaptation-to-mitigation spectrum. This axis ranges from applications focusing purely on adaptation (that is, responding to the effects of climate change to minimize impacts), to applications focusing purely on mitigation (that is, reducing increases in radiative forcing, typically through the reduction of greenhouse gas emissions) to slow or halt climate change.

The first case study focuses on power generation and grid management. The global energy landscape is undergoing a profound transformation, moving away from centralized, fossil fuel-based generation toward a decentralized, dynamic system powered by renewable sources. This transformation is key to successful climate mitigation. However, the intermittent nature of solar and wind power – dependent on weather – poses significant challenges for grid stability and

reliability. AI models are being used for optimizing power generation and enhancing grid management, making it possible to integrate high levels of renewable energy while maintaining a stable, efficient, and resilient grid.⁷

AI models are used to forecast the output of renewable energy sources by leveraging historical weather patterns, satellite imagery, and sensor data from wind turbines or solar panels.⁸ The forecasts range from hours to days in advance. This predictive capability allows grid operators to anticipate fluctuations and adjust other energy sources, such as hydropower or battery storage, to ensure a continuous supply.

In the context of grid management, AI models can perform real-time load balancing by analyzing data from smart meters and other grid sensors to predict electricity demand. This allows the grid to dynamically shift power to where it's most needed and to manage energy storage systems by charging batteries during periods of high renewable output and discharging them during peak demand. This process of demand-response management reduces reliance on "peaker" plants – fossil fuel generators that are typically fired up to meet sudden spikes in demand.

Both applications lend themselves well to the use of AI tools because actual renewable energy production and actual energy demand are monitored throughout the system, making the problem highly tractable. This means that benchmarks to evaluate models are readily available at timescales and at locations that are relevant to the models' performance.

Another area in which short-term forecasts are advantageous is near-term impact assessment and resilience. Climate adaptation is about building resilience to the unavoidable consequences of a changing climate. This resilience spans timescales ranging from hours (for example, during evacuations or emergency aid deployment) to decades (for instance, when planning for sustainable development in the face of a changing climate). Applications of AI models have been especially promising with shorter timescales ranging from hours to seasons. Such applications have focused on anticipating and responding to extreme events such as heatwaves, floods, droughts, and wildfires through early warning systems and dynamic risk assessment for infrastructure and communities.

AI models are being deployed to enable more accurate and timely early warning systems.⁹ Traditional meteorological models, while powerful, are often computationally intensive and struggle to capture local-scale, rapid-onset phenomena. AI, particularly deep-learning models like neural networks, can process vast, multi-modal datasets in near real time to make highly localized predictions. For example, by analyzing satellite imagery, radar data, and ground-based sensor readings, AI can predict the path and intensity of a hurricane, the likely locations of a flash flood, or the spread of a wildfire with greater accuracy than ever before.

This short-term impact assessment goes beyond simple weather forecasting. AI models can integrate meteorological data with other datasets – such as topograph-

ical maps, urban infrastructure layouts, and population density information – to predict the specific impacts of an event. For instance, a model can forecast not only that a river will flood but also which specific streets will be inundated, how many homes are at risk, and where vulnerable populations might be. This information is invaluable for emergency services, allowing them to pre-position resources, issue targeted evacuation orders, and optimize response logistics.

AI tools can also be used to optimize resource allocation for disaster response and recovery.¹⁰ After a flood or earthquake, AI-powered image recognition can quickly analyze satellite or drone imagery to assess damage to buildings and roads. This allows relief agencies to prioritize aid delivery to the most affected areas.

What makes many of these short-term applications highly tractable is the availability of historical data on the precise variables that the AI models are trying to predict. For example, the spread and evolution of historical fires, as captured through satellite imagery, can be used both to train and to evaluate models that aim to predict the evolution of future fires. This means that rich benchmarks are available, and those benchmarks are well matched with the responses that the models aim to predict.

Two aspects make applications in these areas somewhat less tractable than in power generation, however. The first is that many applications related to adaptation focus on predicting and responding to extreme events, which are, by definition, relatively rare. This means that the availability of historical data for training and evaluation can be limited, especially over the relatively brief period during which high-resolution satellite imagery and meteorological data have been available.

The second is that for applications related to disaster response, it is less clear which benchmarks are most appropriate for model evaluation. For example, if an AI model is used to optimize deployment of resources in disaster recovery, how should one evaluate the recommendations of one model versus another, or the overall performance of one specific model? Without access to hundreds (or thousands) of similar events, or the ability to provide counterfactual examples where different approaches were used under the same circumstances, evaluation metrics are difficult to define or quantify. Synthetic data based on simulated events can be used to some extent, and the use of synthetic data makes it possible to generate thousands or even millions of scenarios, but determining whether the synthetic scenarios are sufficiently representative of real-world applications is not trivial.

Contrail (condensation trail) avoidance presents an interesting opportunity on the climate mitigation side of the opportunity space. Contrails are the white line-shaped clouds that are sometimes left in an aircraft's wake. If the ambient air is very dry, these trails dissipate quickly. If the contrails form in regions of ice supersaturation in the upper atmosphere (near the tropopause), they persist and spread. In this case, the small amount of soot and water ice produced by an aircraft acts as condensation nuclei and can draw ten thousand times as much

water vapor out of the atmosphere.¹¹ This causes a thin cirrus cloud that can last for many hours over a wide region.¹²

Contrail cirrus is a notable contributor to climate change because it absorbs outgoing infrared radiation from Earth and reradiates some of that energy back to the planet. In addition to this warming effect, contrail cirrus reflects a small amount of shortwave sunlight back out to space, producing a cooling effect. During the day, the two effects roughly cancel, but at night when there is no shortwave radiation reflection, contrails add net forcing to the climate system.¹³ If one measures climate forcing using the GWP 100 metric (global warming potential over one hundred years), then contrails contribute approximately one-third of the total climate impact of today's aviation.¹⁴

Rerouting air traffic to avoid contrail cirrus is thus a cost-effective method for combating the climate impact of aviation, often requiring only small adjustments to flight altitude or routing that incur minimal fuel penalties.¹⁵ The rerouting is highly effective because approximately 5 percent of flight kilometers produce persistent contrails, and of those, 15 percent produce 80 percent of the positive climate forcing.¹⁶ An AI system could therefore assess the current weather and the flight plan data and predict whether a contrail will form. Such a system can be applied at space-time locations that are likely to produce highly-warming contrails and avoid a large fraction of the climate forcing attributable to contrails.¹⁷

Building such an AI system is highly tractable because we can detect contrails from satellite images.¹⁸ This satellite imagery provides a large, labeled dataset for training machine-learning models. Therefore, the input and output of such a system are dense and well determined, with a clear metric for the AI to use.

Knowing the climate forcing produced by contrails at a given space-time location, as opposed to using long-term averages, would allow us to potentially reduce the number of diversions twofold by focusing on those flights for which the resulting contrails would have the most impact on radiative forcing (that is, warming).¹⁹ However, estimating the contrail cirrus radiative forcing is a much harder problem for AI than simply predicting the formation of contrails because we cannot directly measure the forcing due to a contrail (since we cannot observe the counterfactual atmosphere where the plane did not fly). The current state of the art is to use approximate physical models to estimate the radiative forcing.²⁰ Because of unknowns due to ice microphysics, upper atmosphere humidity, and wind fields, such physical models have large uncertainties. How to apply artificial intelligence in such a scenario is an open research problem.

Quantifying greenhouse gas emissions and sequestration is central to broader efforts in climate mitigation. Indeed, progress toward a net-zero global economy is predicated on the ability to track emissions of greenhouse gases (GHGs) such as carbon dioxide (CO₂) and methane (CH₄) into the atmosphere across

the many sectors of human activity. In parallel, there is a need to track removals of carbon from the atmosphere, whether by the natural components of the earth system (such as plants, soils, and oceans) or by engineered approaches to carbon capture. This is because the natural components of the carbon cycle have been removing approximately half of the carbon that has been emitted through human activity over the past several decades, but the details of where, why, and how this “natural sink” operates are still unclear, as is the degree to which this sink will continue to function as climate change intensifies.²¹

Traditionally, this movement of GHGs into and out of the atmosphere, collectively referred to as “fluxes,” has principally been tracked through one of three approaches. The first is inventories, which rely heavily on generalized activity data and established emission factors. The second is mechanistic models of the natural components of the earth system, which rely on parametric representations of processes like photosynthesis to estimate fluxes. The third is atmospheric inversions, which couple observations of the concentrations of GHGs in the atmosphere with a model representing atmospheric dynamics to trace back the location and magnitude of fluxes. These traditional approaches have strengths and weaknesses, but they are all fundamentally limited by their inability to integrate information across the vast quantity and diversity of atmospheric and Earth surface observations that have become available through major advances in the ability to monitor Earth from space.

Artificial intelligence and machine learning have emerged as pivotal technologies, offering a path toward quantification of all constituent human and natural components of the global carbon budget. This opportunity is thanks to AI’s ability to process vast, heterogeneous datasets and to discover and leverage the interrelationships among variables.

One example of where these tools have already had a large impact is in the development of models that can use small-scale observations of GHG fluxes from trees, plants, and soils, and extrapolate them in space and time to obtain estimates of biospheric activity everywhere, all the time. These tools infer relationships between observations from what are known as eddy covariance flux towers (that can represent fluxes for an area of approximately one square kilometer with meteorological data and satellite observations of Earth’s surface and atmosphere. Once those relationships are captured by the model, meteorological and satellite observations can be used to quantify fluxes where no towers exist.²²

A second example is the use of deep-learning algorithms and convolutional neural networks applied to satellite imagery to enable detection and attribution of point-source and diffuse fugitive plumes, such as those from the oil and gas sector, agriculture, and landfills.²³ These AI systems identify and characterize plumes at a high spatial resolution, significantly reducing the latency of detection and providing actionable data to operators.

The challenge with many of these applications, however, is the dearth of benchmark data to train and evaluate models. This is because fluxes of greenhouse gases cannot be measured directly at most spatial and temporal scales that are directly relevant to climate action. In the case of methane leaks, for example, if the goal is to spot unknown leaks, and if known leaks are fixed, where does one come up with the thousands or millions of example leaks on which to train an AI model? Here again, synthetic data, if carefully designed to mimic real-world situations, can in some cases be used for model development and evaluation. In the case of biospheric GHG fluxes, while AI models can be trained on observations from eddy covariance flux towers, there are only about a thousand such towers globally, with some regions and ecosystems very poorly represented, making it difficult to assess how models perform across vast areas. This presents a fundamental contrast with applications such as weather forecasting, in which tomorrow's weather can be used to assess today's forecast.

Overall there is incredible potential for AI to enable a transition toward substantially more robust emissions measurement, reporting, and verification (MRV) and effective assessment of both nature-based and technological approaches to carbon capture and sequestration in support of climate mitigation. These applications present challenges in terms of their tractability, however, and there is a need to develop sophisticated new approaches for leveraging the data and information that we do have to inform these critical applications.

At the far end of the tractability spectrum lies the problem of using AI for long-term climate projection to support multidecadal planning for climate mitigation and adaptation. This application represents the frontier of intractability because it is fundamentally an out-of-distribution generalization problem, for which the historical record is an increasingly inadequate guide to a future, nonstationary world.²⁴ Purely data-driven AI models, which excel at learning statistical patterns from vast amounts of data, are poor at out-of-distribution extrapolation. AI models may have low skill when confronted with future climate states that involve novel physical dynamics or the crossing of critical thresholds not represented in their training history. To further complicate this, information on historical climate conditions spanning timescales of decades to centuries is severely limited by the relatively short period of the record of observations with high spatial and temporal detail, such as those provided by space-based observations.²⁵

This intractability is rooted in three core properties of the earth system. First is nonstationarity: anthropogenic forcing is actively changing the underlying statistics of the climate. The relationships between variables, such as temperature and precipitation patterns, are not fixed over time, violating the core assumption of many machine-learning algorithms that the training and testing data are drawn from the same distribution.²⁶

The second property is internal climate variability: the climate system possesses natural chaotic fluctuations on different timescales. The atmosphere itself erases information in its initial conditions within roughly fifteen days.²⁷ The entire climate system has multiyear fluctuations, such as the El Niño–Southern Oscillation. The inherent variability of the climate system makes it difficult to isolate the forced “signal” of climate change from the limited historical record. Disentangling this internal variability requires large ensembles of simulations, a task for which AI-based emulators are beginning to show promise.²⁸

Third is the existence of potential tipping points within the climate system. The climate system may contain thresholds or bifurcations at which a small additional forcing can trigger a large, abrupt, and potentially irreversible shift in the state of a subsystem, such as the collapse of an ice sheet, a sudden change in ocean circulation, or the breakdown of the biospheric carbon sinks discussed earlier.²⁹ By definition, such unprecedented events are absent from any training data that one might use to develop models, making them nearly impossible for a purely data-driven model to anticipate. Further, any demonstration of a tipping point in any simulator (either AI- or physics-based) is often very difficult to distinguish from numerical artifacts (or even simple programming bugs!) in the model itself, making such tipping points difficult to validate.

These properties force us to conduct a deeper examination of AI models versus traditional, physics-based earth system models (ESMs). ESMs are often presented as the transparent, physically grounded alternative to “black box” AI. This view, however, is an oversimplification. ESMs are highly complex computational artifacts, comprising millions of lines of code. Crucially, they depend on hundreds of parameterizations – simplified, often empirically derived equations used to represent critical processes like cloud formation and atmospheric turbulence that occur at scales too small to be explicitly resolved by the model’s grid.³⁰ The complex, nonlinear interactions among these parameterized components give rise to emergent behaviors that are not always fully understood or predictable, making the ESM difficult to interpret in its own right. Both ESMs and AI models are complex systems, and the goal should not be to find a perfectly transparent model but to develop a rigorous suite of tools to validate, interrogate, and understand the behaviors and limitations of all our models. This set of tools could be improved by incorporating the latest AI techniques.

The successes and challenges illustrated by the five case studies above hint at how the use of AI is likely to evolve in the future. To anticipate how it will support climate science, climate mitigation, and climate adaptation over the next decade and beyond, we need to look at the cutting edge of today’s artificial intelligence, such as the large language model (LLM)–based systems that can generate complex code or models. One such example is the recent LLM and tree

search system designed to create expert-level scientific software that can be scored using a given quality metric and that searches for code that improves it, including through AI-assisted web or literature searches.³¹ Going forward, AI coding systems will make the creation of models much less expensive than at present, which reduces the need for centralized models or libraries that cover many possible cases.

One of the issues with the current use of machine learning–based models in climate is that they are largely black boxes. Domain experts don’t know the limitations of such models *a priori*. However, if climate models were actually generated by LLM systems, then we could use another LLM to inspect the model code and create hypotheses for why the model behaves the way it does. We could generate new diagnostic or experimental code to test these hypotheses. There are early signs of this behavior in the tree search system, whereby the system automatically generates explanations of why a model had a sudden jump in performance while the LLM is searching for alternatives.³²

Casting forward to 2036, we can imagine the use of such model-generating tools in the climate field. Climate models of all sorts will be built by AIs, not just by researchers in academic departments. The AIs will ensure that the models both reproduce current observational data and obey physical laws and long-term qualitative measures of reasonable models. There may be a shared foundational model that is a coarse but full model of the earth system, vetted by the scientific community.³³ Fine-grained predictions would be built via AIs on top of the shared model. Researchers would then spend most of their time understanding observations and models, rather than writing software, sourcing observations, and contending with the limitations of legacy models. Our hope is that such support from artificial intelligence will accelerate scientific progress in all fields, including models used to support climate action.

Three additional outcomes are likely once such future models are possible. First, the ease of assimilating diverse data into AI-created models will increase the utility of observations and thereby incentivize globally coordinated observational programs. More data will constrain the future models and will rapidly improve them (since building the models will be so much easier). Second, the concept of “data science at the singularity” suggests that rapid scientific progress is enabled by frictionless reproducibility and public metrics.³⁴ If artificial intelligence can easily generate models, they can be more easily shared and measured. This will likely lead to an increase in the rate of progress. Third, rather than replacing scientists, such future models will free researchers to focus on their core scientific expertise and use it to guide the AI to do the grunt work needed to support scientific breakthroughs. In other words, AI coding systems will allow scientists at all levels of seniority – from graduate students to senior faculty – to concentrate on the true creative and analytic work of science, rather than getting bogged down in processing data and developing software.

Looking into the future, we must also recognize that the rapid proliferation of machine learning and artificial intelligence presents a critical paradox for climate science: while AI offers essential tools for climate action, the computational infrastructure supporting it carries a significant, accelerating environmental burden.

The most obvious climate impact of AI stems from the energy required for model training and inference.³⁵ While the computational demands associated with training state-of-the-art LLMs, which often involve billions of parameters, result in substantial electricity consumption, they are dwarfed by the energy used for inference: that is, the energy consumed by queries by the end users.

The rise in the number of queries is driving construction and power demand of global data centers. The International Energy Agency (IEA) has projected that data center electricity consumption could more than double by the end of the 2020s. Since a large portion of this electricity is still currently sourced from fossil fuel-based grids, the result is a potential growth in global GHG emissions. Conversely, this also increases the benefit of decarbonizing electricity production, a key step toward climate mitigation.

Interestingly, however, the use of AI for climate science itself does not put a substantial burden on the climate due to electricity usage. Assuming that scientists are using a tool similar to the tree search system, we can estimate the total energy demand by conducting a back-of-the-envelope calculation based on one hundred thousand atmospheric scientists who are each issuing ten thousand typical Gemini queries a week (assuming that each scientist is calling on the AI coding tool on average once per week).³⁶ At 0.24 watt-hours per Gemini query, the weekly energy burden would be approximately 1.4 megawatts, or roughly the same as twelve hundred average U.S. households.³⁷ The overall carbon burden, assuming the typical U.S. carbon intensity of 384 gCO₂/kWh, would be approximately five thousand metric tons of CO₂ per year. This amount is less than one ten-millionth of total global greenhouse gas emissions.

The convergence of climate urgency and the AI revolution has given us a dual mandate. On the one hand, the next era of climate science, mitigation, and adaptation will be powered by algorithms capable of extracting understanding from the planet's immense, complex, and heterogeneous data streams. Artificial intelligence offers an attractive scalable path toward robust measurement, reporting, and verification; grid stability; and near real-time disaster resilience. On the other hand, this very tool carries a significant and growing environmental footprint.

Moving forward requires a collaborative, humble, and clear-eyed approach. Defining the landscape of climate challenges along the axes of tractability and the adaptation-to-mitigation spectrum can serve as a guide. We must vigorously pursue AI solutions where problems are highly tractable, such as in power grid optimiza-

tion and short-term impact forecasting, and where clear benchmarks and dense data records accelerate progress and yield immediate climate benefits.

However, we must also focus on the “intractable” frontier. To do so, we should not treat physics-based and data-driven models as mutually exclusive. The future lies in fusion: utilizing physics-informed AI and LLM-generated models that accelerate discovery while remaining subject to physical laws and community vetting. Critically, we must address the challenge of scale incompatibility: that is, the mismatch between the scales at which we can measure key climate variables and the scales at which we need models to represent them. The scientific community must collaborate to build the benchmarks and comprehensive observational infrastructure necessary to constrain models.

Ultimately, returning to where this essay began, the goal of constrained optimization is not just to find the best possible solution, but to define the true limits within which action can succeed. By being clear eyed about both the immense potential and the inherent limitations of artificial intelligence, we can successfully navigate the challenges ahead and begin to realize the promise of an Algorithmic Planet working in service of a sustainable future.

ABOUT THE AUTHORS

Anna M. Michalak is Founding Director of the Climate and Resilience Hub at the Carnegie Institution for Science and a Visiting Faculty Researcher at Google Research. Her work focuses on the global carbon cycle, inland and coastal water quality, and climate resilience. She served as Co-Chair of the National Academies for Sciences, Engineering, and Medicine midterm assessment of the decadal survey for Earth observations from space, Chair of the Scientific Advisory Board for the European Integrated Carbon Observation System (ICOS), and lead author of the U.S. Carbon Cycle Science Plan. She has recently published in such journals as *Science*, *Nature*, *Proceedings of the National Academy of Sciences*, *Nature Water*, and *Nature Ecology & Evolution*.

John C. Platt is a Google Fellow and leads the Applied Science branch of Google Research, working at the intersection between computer science and physical or biological science. He was previously Distinguished Scientist at Microsoft Research and Deputy Managing Director of its Redmond Lab, as well as Director of Research at Synaptics. He has published in such journals as *Nature*, *Environmental Research Communications*, and *Energy & Environmental Science*.

ENDNOTES

- ¹ Anna M. Michalak and Peter K. Kitanidis, “A Method for Enforcing Parameter Non-negativity in Bayesian Inverse Problems with an Application to Contaminant Source Identification,” *Water Resources Research* 39 (2) (2003): 1344.
- ² James E. Hansen, Pushker Kharecha, Makiko Sato, et al., “Global Warming Has Accelerated: Are the United Nations and the Public Well-Informed?” *Environment: Science and Policy for Sustainable Development* 67 (1) (2025): 6.
- ³ Intergovernmental Panel on Climate Change, *Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, ed. Core Writing Team, Hoesung Lee, and José Romero (IPCC, 2023), <https://www.ipcc.ch/report/ar6/syr>.
- ⁴ Steven C. Sherwood and Chris E. Forest, “Opinion: Can Uncertainty in Climate Sensitivity Be Narrowed Further?” *Atmospheric Chemistry and Physics* 24 (4) (2024): 2679–2686, <https://doi.org/10.5194/acp-24-2679-2024>.
- ⁵ Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, et al., “Learning Skillful Medium-Range Global Weather Forecasting,” *Science* 382 (6677) (2023): 1416–1421; Kaifeng Bi, Lingxi Xie, Hengheng Zhang, et al., “Accurate Medium-Range Global Weather Forecasting with 3D Neural Networks,” *Nature* 619 (2023): 533–538; and Ravi Aggarwal, Viknesh Sounderajah, Guy Martin, et al., “Diagnostic Accuracy of Deep Learning in Medical Imaging: A Systematic Review and Meta-Analysis,” *NPJ Digital Medicine* 4 (2021): 65.
- ⁶ Nestor Maslej, Loredana Fattorini, Raymond Perrault, et al., *The AI Index 2025 Annual Report* (AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, 2025), https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf.
- ⁷ Keith J. Benes, Joshua E. Porterfield, Charles Yang, et al., *AI for Energy: Opportunities for a Modern Grid and Clean Energy Economy* (U.S. Department of Energy, 2024), https://www.energy.gov/sites/default/files/2024-04/AI%20EO%20Report%20Section%205.2g%281%29_043024.pdf.
- ⁸ Seong Lok Choi, Rishabh Jain, Cong Feng, et al., *Generative AI for Power Grid Operations* (National Renewable Energy Laboratory, U.S. Department of Energy Office of Energy Efficiency and Renewable Energy, 2024), <https://www.osti.gov/biblio/2477920>.
- ⁹ Markus Reichstein, Vitus Benson, Jan Blunk, et al., “Early Warning of Complex Climate Risk with Integrated Artificial Intelligence,” *Nature Communications* 16 (2025): 2564.
- ¹⁰ Patrick S. Roberts, “How AI Is Changing Our Approach to Disasters,” RAND Corporation, August 27, 2025.
- ¹¹ David Atlas, Zhien Wang, and David P. Duda, “Contrails to Cirrus—Morphology, Microphysics, and Radiative Properties,” *Journal of Applied Meteorology and Climatology* 45 (1) (2006): 5–19.
- ¹² Patrick Minnis, David F. Young, Donald P. Garber, et al., “Transformation of Contrails into Cirrus during SUCCESS,” *Geophysical Research Letters* 25 (8) (1998): 1157–1160.
- ¹³ Nicola Stuber, Piers Forster, Gaby Rädel, and Keith Shine, “The Importance of the Diurnal and Annual Cycle of Air Traffic for Contrail Radiative Forcing,” *Nature* 441 (2006): 864–867.
- ¹⁴ Alaa Al Khourdajie, Renée van Diemen, William F. Lamb, et al., “Annex II: Definitions, Units and Conventions,” in *Climate Change 2022: Mitigation of Climate Change. Contribution*

- of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, ed. Priyadarshi R. Shukla, Jim Skea, Andy Reisinger, et al. (Cambridge University Press, 2022), 1821–1840, <https://doi.org/10.2172/1973107>; and David S. Lee, David W. Fahey, Agnieszka Skowron, et al., “The Contribution of Global Aviation to Anthropogenic Climate Forcing for 2000 to 2018,” *Atmospheric Environment* 244 (2021): 117834.
- ¹⁵ Roger Teoh, Ulrich Schumann, and Marc E. J. Stettler, “Beyond Contrail Avoidance: Efficacy of Flight Altitude Changes to Minimise Contrail Climate Forcing,” *Aerospace* 7 (9) (2020): 121.
- ¹⁶ Roger Teoh, Zebediah Engberg, Ulrich Schumann, et al., “Global Aviation Contrail Climate Effects from 2019 to 2021,” *Atmospheric Chemistry and Physics* 24 (10) (2024): 6071; and John C. Platt, Marc L. Shapiro, Zebediah Engberg, et al., “The Effect of Uncertainty in Humidity and Model Parameters on the Prediction of Contrail Energy Forcing,” *Environmental Research Communications* 6 (9) (2024): 095015.
- ¹⁷ Aaron Sonabend-W, Carl Elkin, Thomas Dean, et al., “Feasibility Test of Per-Flight Contrail Avoidance in Commercial Aviation,” *Communications Engineering* 3 (2024): 184.
- ¹⁸ Scott Geraedts, Erica Brand, Thomas R. Dean, et al., “A Scalable System to Measure Contrail Formation on a Per-Flight Basis,” *Environmental Research Communications* 6 (2024): 015008.
- ¹⁹ Platt, Shapiro, Engberg, et al., “The Effect of Uncertainty in Humidity and Model Parameters on the Prediction of Contrail Energy Forcing.”
- ²⁰ Ulrich Schumann, Bernhard Mayer, Kaspar Graf, and Hermann Mannstein, “A Parametric Radiative Forcing Model for Contrail Cirrus,” *Journal of Applied Meteorology and Climatology* 51 (7) (2012): 1391–1406; and Chien-Cheng Chen and Andrew Gettelman, “Simulated Radiative Forcing from Contrails and Contrail Cirrus,” *Atmospheric Chemistry and Physics* 13 (24) (2013): 12525–12536.
- ²¹ Paola A. Arias, Nicolas Bellouin, Erika Coppola, et al., “Technical Summary,” in *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, ed. Valérie Masson-Delmotte, Pan-mao Zhai, Anna Pirani, et al. (Cambridge University Press, 2021), 33–144.
- ²² Jacob A. Nelson, Sophia Walther, Fabian Gans, et al., “X-BASE: The First Terrestrial Carbon and Water Flux Products from an Extended Data-Driven Scaling Framework, FLUXCOM-X,” *Biogeosciences* 21 (22) (2024): 5079–5115.
- ²³ Berend J. Schuit, Joannes D. Maasakkers, Pieter Bijl, et al., “Automated Detection and Monitoring of Methane Super-Emitters Using Satellite Data,” *Atmospheric Chemistry and Physics* 23 (16) (2023): 9071–9098, <https://doi.org/10.5194/acp-23-9071-2023>; and Anna Vaughan, Gonzalo Mateo-García, Luis Gómez-Chova, et al., “CH4Net: A Deep Learning Model for Monitoring Methane Super-Emitters with Sentinel-2 Imagery,” *Atmospheric Measurement Techniques* 17 (9) (2024): 2583–2593.
- ²⁴ Stefano Materia, Lluís Palma García, Chiem van Straaten, et al., “Artificial Intelligence for Climate Prediction of Extremes: State of the Art, Challenges, and Future Perspectives,” *WIREs Climate Change* 15 (6) (2024): e914.
- ²⁵ Jun Yang, Peng Gong, Rong Fu, et al., “The Role of Satellite Remote Sensing in Climate Change Studies,” *Nature Climate Change* 3 (10) (2013): 875; and Annalisa Bracco, Julien Brajard, Henk A. Dijkstra, et al., “Machine Learning for the Physics of Climate,” *Nature Reviews Physics* 7 (1) (2025): 6–20.

- ²⁶ Keith W. Dixon, John R. Lanzante, Mary Jo Nath, et al., “Evaluating the Stationarity Assumption in Statistically Downscaled Climate Projections: Is Past Performance an Indicator of Future Results?” *Climatic Change* 135 (3–4) (2016): 395–408.
- ²⁷ Fuqing Zhang, Y. Qiang Sun, Linus Magnusson, et al., “What Is the Predictability Limit of Midlatitude Weather?” *Journal of the Atmospheric Sciences* 76 (4) (2019): 1077–1091.
- ²⁸ Veronika Eyring, William D. Collins, Pierre Gentine, et al., “Pushing the Frontiers in Climate Modelling and Analysis with Machine Learning,” *Nature Climate Change* 14 (9) (2024): 916–928.
- ²⁹ Sebastian Bathiany, Henk Dijkstra, Michel Crucifix, et al., “Beyond Bifurcation: Using Complex Models to Understand and Predict Abrupt Climate Change,” *Dynamics and Statistics of the Climate System* 1 (1) (2016).
- ³⁰ Norman McFarlane, “Parameterizations: Representing Key Processes in Climate Models Without Resolving Them,” *WIREs Climate Change* 2 (4) (2011): 482–497.
- ³¹ Eser Aygün, Anastasiya Belyaeva, Gheorghe Comanici, et al., “An AI System to Help Scientists Write Expert-Level Empirical Software,” arXiv (2025), <https://doi.org/10.48550/arXiv.2509.06503>.
- ³² Ibid.
- ³³ Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, et al., “ClimaX: A Foundation Model for Weather and Climate,” in *Volume 202: Proceedings of the 40th International Conference on Machine Learning*, ed. Andreas Krause, Emma Brunskill, Kyunghyun Cho, et al. (MLR Press, 2023), 26032–26056, <https://proceedings.mlr.press/v202>; and Aaron Bell, Amit Aides, Amr Helmy, et al., “Earth AI: Unlocking Geospatial Insights with Foundation Models and Cross-Modal Reasoning,” arXiv (2025), <https://doi.org/10.48550/arXiv.2510.18318>.
- ³⁴ David Donoho, “Data Science at the Singularity,” *Harvard Data Science Review* 6 (1) (2024).
- ³⁵ International Energy Agency, *Energy and AI* (International Energy Agency, 2025), <https://www.iea.org/reports/energy-and-ai>.
- ³⁶ Aygün, Belyaeva, Comanici, et al., “An AI System to Help Scientists Write Expert-Level Empirical Software.”
- ³⁷ Cooper Elsworth, Keguo Huang, David Patterson, et al., “Measuring the Environmental Impact of Delivering AI at Google Scale,” arXiv (2025), <https://doi.org/10.48550/arXiv.2508.15734>.

AI Reaches for the Stars

Stella S. R. Offner

Artificial intelligence is rapidly being integrated into a variety of tasks in astronomical research, from model fitting to anomaly detection. While AI offers huge potential to accelerate discoveries about the universe, its current capabilities and trustworthiness fall short. I discuss the road to full AI integration into astronomical workflows: why the field of astronomy is a particularly fertile ground for AI development, the potential revolutionary gains possible for scientific analysis and discovery, how AI developments within astronomy can accelerate AI advancements more broadly, and the current roadblocks to fully AI-integrated astronomy.

In recent years, progress in artificial intelligence has exploded, achieving previously unimaginable advances in image and language processing. Meanwhile, engagement with AI methods within scientific communities like astronomy has grown exponentially on two fronts. Over the past decade, formal use by researchers to apply AI methods to data analysis, including to classify objects, estimate parameters, and identify outliers, has increased tenfold.¹ Meanwhile, the recent rise of powerful public large language models (LLMs) such as OpenAI’s ChatGPT, Anthropic’s Claude, and Deepseek has sparked a quiet revolution in the adoption of AI to summarize research papers, draft and revise text, and write code. A 2024 survey of nearly five hundred astronomers by the American Astronomical Society found that 33 percent of respondents had used AI as a writing assistant and 48 percent had used AI to program.² AI is rapidly becoming not only common within but also an indispensable component of research workflows.

While embrace of AI by the full astronomy community remains somewhat distant, the future is clear. AI will evolve beyond being just another useful tool to become an intelligent assistant or “copilot” that is fully integrated in all aspects of the scientific process: hypothesis generation, information gathering via telescope observations or archival database queries, data processing and analysis, and data presentation in journals. The resulting AI boost to productivity and efficiency has the power to accelerate the research process; projects that currently take months or years will be completed in days.

My own journey into AI began in the early 2010s via an interest in image classification. Astronomy, as a fundamentally observational science, has a unique relationship with image processing. Preeminent telescopes like the Hubble Space Telescope (HST)

and Spitzer Space Telescope have generated huge mosaics of the sky, capturing the ethereal beauty of star clusters, galaxies, and nebulae. The Spitzer Galactic Plane Survey panned across the midline of our Milky Way Galaxy to produce an enormous image of 390,000 x 6,000 pixels, peppered with supernovae shells, wispy irradiated dust clouds, and winding dark lanes where stellar nurseries enshroud newborn stars.³ These complex details are readily identified by eye, but classical automated pattern searches fare poorly, since there are many overlapping, faint features, and even bright sources are complex and heterogeneous. This astronomy data challenge inspired many citizen science projects, including Galaxy Zoo and the Milky Way Project, which, by the mid-2010s, had engaged hundreds of thousands of volunteers to visually identify and classify tens of millions of celestial objects.⁴

My own research investigates the complex energy signatures that young stars produce in the form of high-velocity jets, wind-blown bubbles, radiatively driven shells, and supernovae. These features, captured in high-definition images taken by the Hubble, Spitzer, and James Webb Space Telescopes, shape the stellar birth environment, including how efficiently stars form and when star formation stops. As a computational astrophysicist, my primary research instruments are supercomputers rather than telescopes. By the 2010s, numerical simulations of star formation had reached a volume, level of complexity, and accuracy to make ideal training sets for AI models.

Unlike human visual identifications, the outputs of computer simulations offer reliable, labeled data, while theoretically circumventing human biases. My foray into image analysis began with *random forests*, an old but robust approach to classification that is based on decision trees.⁵ Although this method performed comparably to visual identification, it was not flexible enough to automatically identify complex stellar feedback features.⁶ I turned to convolutional neural networks (CNNs), ultimately adopting an architecture able to identify structures via both spatial and wavelength correlations.⁷ My research group developed the Convolutional Approach to Shell Identification in three dimensions (CASI-3D) to rapidly comb through large observational datasets and find stellar feedback at the level of individual voxels, contouring features to provide three-dimensional maps of their locations and properties.

Today, due to the complexity of astronomy data, many analyses continue to be conducted “by eye.” Galaxy Zoo is still going strong! However, a broad range of astronomy tasks, from outlier detection to model fitting, are being automated through AI methods.

So what exactly does it mean to use AI in astronomy research? The rapid growth of AI capabilities and use have generated a plethora of definitions that reflect AI application in different contexts. Here, I define AI in the broadest sense: intelligent machines that learn to complete tasks without being explicitly programmed.⁸ *AI machines* could refer to intelligent robots, which learn from their

environment; however, in the context of science, AI refers to intelligent *algorithms*, which are trained to perform tasks without being programmed with explicit rules. Just as ChatGPT learned to write poetry from reading large numbers of poems, sophisticated astronomy AI models trained on massive observational datasets learn to differentiate galaxies, stars, and accreting blackholes without being given instructions for how to identify these sources.

AI algorithms, like humans, have the ability to learn symbolic representations of concepts.⁹ Some AI methods – namely, those architectures based on artificial neural networks, like CASI-3D – are direct analogs to the human brain, updating according to new input information and responding to positive or negative feedback (rewards or punishments) from the environment. AI methods also include LLMs, as well as vision transformers and diffusion models, like the algorithms underlying DALL-E, a commercial image generation model. Machine learning (ML), the subfield of AI commonly applied in the context of scientific data analysis and modeling, refers to algorithms trained to make data-driven decisions and predictions.¹⁰ Common ML tasks include classification (as with random forests) and parameter estimation (regression).

To understand the full potential for AI within astronomy today and why astronomy is a unique sandbox for AI requires an appreciation for the history of astronomical data and collaboration. The astronomy ecosystem is uniquely conducive to rapid AI development due to the standardization of practices that promote frictionless reproducibility: freely shared data, open-source code, and driving challenges – three conditions that accelerate data-driven technological advances.¹¹ This inclusive research environment was cultivated through decades of internationally coordinated efforts to digitize observational data, standardize data formats, and promote free open-source tools.

By the mid-1800s, a large number of astronomical observatories were operating around the world, recording images of the night sky on photographic plates. These plates, which were made with glass or metal and covered with a light-sensitive coating, could reach sizes of 30 cm x 30 cm; they were both a poorly transportable data record and owned by the observer, who had little incentive to loan these data to other researchers. Photographic film began to supplant plates in the 1950s, although data distribution and sharing remained challenging. In the 1960s, advancements in computing and the rapidly growing wealth of data incentivized efforts to digitize telescope observations.¹² Early preparations for the design and launch of the Hubble Space Telescope accelerated these efforts and, by the mid-1990s, astronomy had fully entered the digital era.¹³ A variety of subsequent efforts have scanned and digitized photographic plates, further expanding the public archive of astronomical data.¹⁴

Despite the growing collection of digital data, access was also inhibited by heterogeneous data storage formats, the challenges of manipulating the stored data

without a shared software toolkit, and the need to combine observations across the electromagnetic spectrum from different facilities, each with bespoke data recording practices. In the 1970s, this state of affairs motivated the development of a standard data storage format. In 1982, the International Astronomical Union urged the adoption of the Flexible Image Transport System (FITS), a syntax for recording, sharing, and archiving data.¹⁵ More than forty years later, FITS remains the default global format for storage of all astronomical data products.

In the absence of open-source software and public data repositories, the adoption of the FITS format was not sufficient to produce universal data sharing. Over the next two decades, community pressure and funding from NASA and the National Science Foundation (NSF) gradually promoted the creation of open-source community tools, leveling the field for data access and exploration. These developments were further accelerated by large astronomical survey projects, including those led by the HST, which sparked the creation of extensive digital archives or “virtual observatories” that further accelerated the democratization of astronomy data access. Every year since 2003, more than half of the papers published with HST data use archival data rather than new observations.¹⁶

In today’s competitive AI landscape, in which datasets are monetizable, collaboration and free exchange within astronomy are further facilitated by the low economic value of astronomical data. The late astronomy enthusiast and computer scientist James Gray joked that he “liked working with astronomers because their data is worthless.”¹⁷ In other words, because astronomy data are free from the privacy, ethical, legal, and national security issues of other datasets, astronomy offers an AI playground. This culture of open data is reinforced by decades of national and international funding to support transnational astronomy infrastructure and research, which requires open data access and community sharing. Federal support has historically been buoyed by an enthusiastic public, inspired by the wonder of the night sky and universal appeal of the question: How did the universe come to be?

Contemporaneous with the recent revolution in AI, progress in observational capabilities and computing has opened new frontiers for astronomical data analysis. Astronomy datasets are increasingly large, complex, and heterogeneous (“multimodal”), creating fresh barriers to data access and exploration. Emerging AI methods offer enormous potential to address inefficiencies and bottlenecks that touch all aspects of the astronomy research workflow.

Hypothesis Generation: Sparking New Research Directions

The rapid advancement of LLMs promises a radically new approach to search and synthesize online knowledge, including scientific journal articles, catalogs of archival data, and even databases of future observing campaigns (such as abstracts

and specifications of accepted telescope proposals, which are publicly available in astronomy). Due to their extensive training data, general LLMs like GPT-5 can expertly answer high-level astronomy questions. In tandem, there is increasing interest in specialized domain-specific LLMs. Focused models tuned for narrower application are potentially computationally cheaper and have better-defined and more-transparent inputs. In 2023, the first open-source astronomy-domain LLM for literature review, AstroLLaMa, was trained on three hundred thousand paper abstracts.¹⁸ While more prototype than assistant, AstroLLaMa was shortly followed by AstroLLaMa-2 and Pathfinder, which included larger training sets containing partial or full journal articles as well as various journal metadata.¹⁹ These larger astronomy models perform significantly better on the search and synthesis of astronomy journal archives. To date, the performance of the models emerging from academic research lags behind that of state-of-the-art general, commercial models like GPT-4o (*o* for *omni*) and Claude.²⁰

Nonetheless, value is added by the development of domain-specific features. Pathfinder is accompanied by a tool to visualize the astronomy research “landscape.” The topology of this literature map allows researchers to visualize topic relationships and quickly identify areas of missing work or emergent interest. Digestion of the full corpora of knowledge enables such models to suggest novel research directions.²¹ However, this capability remains an active area of research, as the promise of LLM-suggested research directions still falls short compared with those proposed by expert researchers.²² Despite the current deficits in LLM complex reasoning, there is cause for optimism: the steep upward trajectory of AI methodology suggests LLM ideation will soon equal or surpass human hypothesis generation, thereby advancing research frontiers in new and unexpected directions.

Automating Telescope Operations: Collecting & Wrangling Big Data

Multinational, multibillion dollar observatories, like the recently launched James Webb Space Telescope, are the flagship research laboratories that drive astronomical discoveries. The capabilities and power of such facilities have been steadily accelerated by computing and technological innovations. Next-generation facilities, such as the ALMA (Atacama Large Millimeter/submillimeter Array) Wideband Sensitivity Upgrade, next-generation Very Large Array, Square Kilometre Array, and U.S./European Extremely Large Telescopes, will produce a firehose of rich data, far exceeding current processing capabilities.

Despite advancements in detector/instrument capabilities, the data-processing pipeline – from light detection to user delivery – is not fully automated. Data quality assurance, which includes vetting the impact of changing weather conditions and occurrence of instrumental artifacts, is substantially human-limited. The current reliance on human effort by domain experts will not scale with the data volume

expected from near- and long-term planned facilities without a paradigm shift in methodology.

Upon delivery to the researcher, the large volume (petabyte-scale) of data products hinders efficient scientific analysis, even in the context of archival research. For example, a single observation using the next-generation ALMA Wideband Sensitivity Upgrade detector will increase dataset sizes seventy-five-fold, recording light information across 1.2 million wavelength channels. In under a decade, observing projects for individual investigators will regularly generate *trillion* voxel terabyte-scale datasets, exceeding the capacity of typical PCs. Moreover, such cubes are often noisy and signal-sparse and thus challenging to parse for relevant features, especially for researchers without easy access to high-performance computing resources. Such data may have thousands of atomic and molecular features, many of which will be faint or have a structure that is not known *a priori*.²³ Emerging AI techniques for online analysis, outlier detection, and signal processing are critical to streamline data workflows by efficiently extracting salient features that indicate operational anomalies in the data collection or that are scientifically interesting for further study.

Accelerating Simulations: Fast Calculations of Astrophysical Systems

Machine-learning methods offer enormous potential to accelerate the modeling of astrophysical systems. Supercomputers are the laboratories of astrophysics research. Since it is not possible to travel outside our solar system to collect samples or to recreate the extreme conditions of space in laboratories on Earth, supercomputers fill a vital experimental niche, enabling researchers to solve the equations of the universe and explore how different conditions and physical laws shape our reality.

Simulations running on the world's largest supercomputers have been used to calculate how the early universe expanded, model the evolution of galaxies, and study the formation of stars and planets.²⁴ Solving the fundamental equations of nature, however, is not cheap: these calculations typically require many millions of processor hours. Machine-learning methods offer several promising shortcuts to the traditional approach that directly solves the underlying physics equations. Neural networks may contain billions of tunable parameters, which provide the flexibility to describe complex physical systems. Given sufficient input training data, neural networks act as extremely high-dimensional fitting functions, able to represent complex, nonlinear dynamical behavior. Meanwhile, regression-based approaches aim to discover new, more-efficient expressions of the underlying physical laws, thereby allowing much faster calculations. AI methods can also be used to map the problem into a different mathematical space of lower or higher dimension, leading to a more efficient representation that is more computationally tractable.

AI acceleration is not without pitfalls, however. Accuracy depends on vast amounts of high-quality, labeled training data, which requires running large, very

expensive suites of simulations. This catch-22 only works in favor of AI applications if the simulations and data are available *a priori* and do not exceed the cost of calculating the solution directly. This points to training foundation models, which are sufficiently powerful to generalize beyond some narrow scope. In addition, many proof-of-concept studies demonstrate accuracy using idealized, low-dimensional problems (such as two-dimensional, smooth fluid flows and swinging pendulums). Architectures that perform well on simple benchmarks may not achieve the needed accuracy to model the high dynamic range and extreme conditions of astrophysical systems.²⁵ In addition, processes like star formation and galaxy evolution include phenomena that span timescales from days to billions of years. Classical techniques to circumvent short timescales, such as travel times imposed by the speed of light, produce significant spatial coupling across the problem domain, incurring large communication costs. Promising AI applications for astrophysical modeling involve replacing the most costly and modular components of direct simulations, like stellar evolution, gravitational dynamics, and chemical reactions.²⁶ Molecule formation in dense gas, or “astrochemistry,” is a prime example, since chemical reactions may occur in seconds, but the resulting distribution of species is a crucial input to the gas temperature, which coevolves with other processes that span thousands of years. The dynamics of few-body stellar/planetary interactions is another promising application, since the outcomes may be represented by statistical sampling, which is otherwise costly.²⁷

Current research has only scratched the surface of the full potential for modeling astrophysical systems. Next-generation AI models able to generalize to a broad range of physical conditions – while performing with the fidelity of classical models – have the power to transform how we model the universe.

Supercharging Data Analysis: Fitting, Explaining & Predicting Physical Parameters

Science is fundamentally grounded in statistics: all measurements have measurement errors. Consequently, comparing models to data requires a rigorous estimation of measurement confidence. The degree of the uncertainty in a given observation makes the difference between, say, an accelerating universe and a decelerating universe.²⁸ Testing and validating theoretical models requires a detailed model not only for a given parameter but for the possibility or *likelihood* of obtaining different results.

While the statistical frameworks concerned with estimating likelihoods predates AI/ML, AI offers fresh, efficient approaches for estimating measurement expectations. Estimating likelihoods for astronomical parameters is particularly challenging, since observations suffer from the fundamental limitation that physical properties – like density, temperature, or mass – cannot be directly measured but instead must be *inferred* from the characteristics of the observed light. Mapping

between observations and theoretical predictions requires assuming a physical model for the system that itself contains uncertainties. This “inverse” problem is common throughout science, but it is particularly vexing in astronomy, where physical scales span many orders of magnitude, spatial and velocity information are missing, evolution occurs on timescales dwarfing human lifetimes, and some of the key model parameters concern substances – dark energy and dark matter – that do not even emit light.

AI techniques combined with detailed simulations of physical systems, or “simulation-based inference,” allow complex likelihoods to be quantitatively estimated even when the underlying system properties are uncertain. Simulation-based inference techniques have been particularly powerful in driving progress in the field of cosmology, where AI methods combined with thousands of simulations of the universe and observational surveys encompassing millions of galaxies are able to tightly constrain the nature of dark matter and the expansion rate of the universe.²⁹ AI-based inferences thereby allow astronomers to confidently rule out sets of theoretical models more quickly and more robustly than with classical statistics.

AI successes in cosmology are partially owed to the relatively simple initial conditions of the early universe, which can be described by a small number of parameters that are (relatively) easy to sample and model. The situation becomes less tractable when the physical conditions of a given system are poorly constrained and no first-principles mathematical description exists, as in the subfields of star and planet formation. However, even in the simplest physical problems, the universe is complex; simulations, which are critical for AI inference, remain approximations of the physical world, or to quote the late statistician George Box, “all models are wrong.” As computing power continues to grow and simulations become even larger and more realistic, AI advances are needed to provide better metrics to quantify how simulations diverge from reality.

But even imperfect models advance our understanding of the true nature of the universe. Mismatches between models and measurements reveal our ignorance and misconceptions, with AI pointing the way to more accurate representations of our reality. In other words, while all models are wrong, even flawed models are useful.

While astronomy research clearly benefits from foundational advances in AI, the astronomy domain in turn offers enormous potential to spark AI innovation. Continued AI growth demands rich, accessible data, substantive novel challenges, and expert talent, all of which astronomy provides in abundance.

Astronomical data, unlike data in many scientific fields, are public and plentiful. Flagship observatories produce extremely high-dimensional and low signal-to-noise data, providing large and rich datasets able to power AI development. These astronomical data are inherently complex (multimodal), presenting a challenging application space that offers extensions to a broad range of other fields.

In contrast, public data that are siloed by facility, researcher, or application due to heterogenous storage conventions are not truly accessible. Improper or incomplete data documentation creates further impediments to use. Efficiently training AI models requires standardized data annotated with clear, complete metadata. The field of astronomy, having globally adopted a uniform data format (FITS) decades ago that encompasses multiwavelength observations from facilities around the world, is significantly ahead of the curve compared with other scientific disciplines.

In addition, the nonpropriety nature of astronomy data allows AI researchers to experiment and develop new capabilities without cumbersome legal, ethical, and financial restrictions. Because astronomy research occurs in an open-source, collaborative ecosystem, it provides unparalleled opportunities to develop and test AI methods in a “safe” public domain.

Beyond plentiful, rich, interoperable data, human capital is an essential fuel powering the AI revolution. AI advances depend critically on the availability of an extensive AI-skilled workforce. Astronomy, through the wonder of the night sky, is a natural gateway into science, technology, engineering, and math (STEM) fields. Astrophysics research, which engages a broad range of undergraduate students, graduate students, and post-PhD researchers, provides intensive training in wrangling, modeling, and predicting on large, complex datasets. Consequently, academia is an important pipeline of talent for the private sector; more than half of junior astronomy researchers ultimately transition to industry jobs.³⁰ The universal appeal of astronomy and its unique research ecosystem are a springboard to accelerate the training of the future AI workforce.³¹

In recognition of these synergies, in 2024, the National Science Foundation and Simons Foundation partnered to fund two five-year AI institutes focused on astronomy: the NSF-Simons AI Institute for Cosmic Origins (CosmicAI) and the National AI Institute for the Sky (SkAI). These two institutes expand the National AI Institutes program, which has funded twenty-nine institutes to date “to foster long-term, fundamental research in AI . . . and reinforce the foundation of technical leadership in AI.”³² These new AI institutes will confront pressing challenges in astronomical research, while expanding the frontiers of AI capabilities and building infrastructure to train the next-generation workforce.

Despite the tremendous promise of AI approaches, the road to fully AI-integrated scientific workflows faces formidable obstacles. In addition to various technical challenges, progress is stymied by a lack of educational infrastructure, the inherent interdisciplinary nature of AI application, and a variety of sociological factors, including significant community skepticism, all of which have hampered AI adoption. For example, despite the widespread public embrace of LLMs as writing assistants, in a 2024 survey, 78 percent of astronomers reported being either somewhat or very concerned about using AI to write scientific papers,

while 50 percent were either somewhat or very concerned about using AI to analyze data.³³ This trepidation is rooted in (legitimate) apprehension about errors produced by AI hallucinations and the often inscrutable nature of AI predictions, concerns that are buoyed by continuing deficits in AI literacy, common among even the highly educated academic workforce.

Despite the rapid growth of AI within the public and private sectors, formal AI training has lagged behind the rising demand for expertise. Moreover, there is little advanced training within the context of AI in science.³⁴ Interested science students must seek out classes in other departments, such as in computer science, math, or data science. However, these courses are often difficult for nonmajor students to access, tend to focus on ML theory over application, and rarely include scientific, much less astrophysical, examples.

Developing AI educational resources and academic infrastructure to train a next-generation AI-skilled workforce is a central mandate of the NSF AI Institute program. Recent steps to expand scientifically relevant AI training include the NSF-Simons CosmicAI–sponsored Graduate Certificate of AI and ML (CAIML) program, which offers low-cost, online graduate courses in fundamental AI methods, including an astrophysical data application course.³⁵ Online programs like CAIML help bridge educational gaps while traditional academic undergraduate and graduate programs in the physical sciences more gradually incorporate domain-specific AI training within their standard curricula – a process that can take years.

Meanwhile, technological challenges remain at the interface of AI and scientific research. Black boxes are fundamentally anathema to science; AI adoption goes hand in hand with AI *trustworthiness*. While scientists are among the most trusted public figures, with 75 percent of international survey respondents trusting the scientific method, significant wariness about AI in science remains.³⁶ High-confidence scientific results have several key attributes, which are often missing in AI solutions. A trustworthy result is *reproducible*, meaning that it is consistently verifiable by independent observers. Scientific measurements must also be *interpretable*; *why* the outcome occurs must be predictable according to a series of logical arguments. Trust also requires *explainability*, meaning that the measurement is embedded within a broader, coherent narrative. For example, measurements of planetary motions around the Sun are *reproducible*. Kepler’s laws of planetary motion provide an *interpretable* framework for the planet positions, while the Newtonian laws of physics *explain* the behavior of planets in the context of the cosmos.

Among the sciences, sociological factors provide an additional confidence boost for astronomy research. Astronomy is an old science, accessible to any with the patience and curiosity to monitor the night sky by eye or with a small backyard telescope. Meanwhile, astronomy research is almost entirely federally funded, subject to little external financial pressure or corporate influence – another perk of such “worthless” data.

The rigorous mathematics underlying astrophysical theories provides a natural framework for reproducibility, interpretability, and explainability. While AI models also rest on a firm mathematical foundation of statistics and linear algebra, they trade simplicity for flexibility. AI models, like the complex workings of the human brain, violate Occam's razor: the simplest AI model is not necessarily the best one. Advances in mathematical theory are needed to provide explainability for high-parameter AI models, with implications for understanding the complex biology of human thought. In the meantime, while AI technologies hold vast promise to accelerate astronomical discoveries, social and cultural factors resist change; in other words, some of the fault is not in our stars but in ourselves.

Just as the first programmable computers in the 1940s sparked a renaissance that ultimately revolutionized how we do science, AI advancements are rapidly broadening the scope of the possible. In the near future, astronomy researchers will be able to conceptualize, define, and execute research projects using trustworthy, efficient, interpretable, and robust AI methods. Astronomers will be able to interact with trustworthy LLMs that are aware of the full canon of astronomy data and research. Fast, robust AI methods will expedite data collection and modeling, while interpretable AI techniques will allow insightful and meaningful conclusions from the analysis.

Prior innovations in astronomical imaging have seeded practical advancements in diverse areas of society.³⁷ The development of adaptive optics systems to confront the distorting effects of turbulence in Earth's atmosphere led to improvements in retinal imaging and laser eye surgery.³⁸ Astronomers were the first to recognize the exciting scientific possibilities of charge-coupled devices (CCDs) – chips that are indispensable components of today's cell phones – and the drive to capture larger, higher-resolution images of the cosmos accelerated CCD technology.³⁹ AI advancements, sparked by our innate human desire to explore the cosmos, offer similar potential for technology transfer.

Despite this bright future, current astronomy AI applications primarily accelerate rather than revolutionize research. To date, there is no clear astronomy analog to the game-changing protein-folding AlphaFold AI model. On the horizon, foundation models trained on the massive next-generation observational and simulation datasets hold vast potential to induce a phase change in research workflows, not only sparking discoveries but changing how research is conducted. However, more work is needed to incorporate and provide uncertainties in model predictions, which are a critical aspect of advancing scientific progress. In the distant future, foundation model-enabled rapid source detection and AI-driven decision-making coupled to telescope infrastructure will be able to seamlessly integrate astronomical datasets, telescope operations, and computational data analysis. This future "AI astronomer" will sort through massive astronomical archives to form and explore novel hypotheses by peering into the depths of the universe through AI-optimized telescope eyes.

Any transformative innovation inevitably raises the question of human value: Will AI astronomers replace human astronomers? The history of technological progression suggests that AI will indeed change how new discoveries are made, demanding new skill sets and workflows. But human astronomers will remain essential. Future scientific progress will be catalyzed by increasingly integrated human-computer collaborations. AI will amplify humanity's reach toward the stars, copiloting the next generation of astronomical discoveries: galaxies forming during the cosmic dawn of the universe, the gravitational chirps of merging black holes, exotic worlds circling distant suns, and even, possibly, the holy grail of astronomical searches, the elusive signals of life beyond our solar system.

ABOUT THE AUTHOR

Stella S. R. Offner is Professor of Astronomy at the University of Texas at Austin. She is also the Co-Director of the Center for Scientific Machine Learning at the Oden Institute for Computational Engineering and Sciences and is the PI and Director of the NSF-Simons AI Institute for Cosmic Origins. She has recently published in such journals as *Nature Astronomy*, *The Astrophysical Journal*, and *Astronomy & Astrophysics*.

ENDNOTES

- ¹ Daniela Huppenkothen, Michelle Ntampaka, Matthew Ho, et al., "Constructing Impactful Machine Learning Research for Astronomy: Best Practices for Researchers and Reviewers," arXiv (2023), Figure 1, <https://doi.org/10.48550/arXiv.2310.12528>.
- ² L. Prato and B. Gaensler, private communication, 2025.
- ³ "GLIMPSE (Galactic Legacy Infrared Mid-Plane Survey Extraordinaire)," Spitzer Space Telescope, <http://ftp.astro.wisc.edu/glimpse> (accessed December 15, 2025).
- ⁴ Galaxy Zoo, <https://www.zooniverse.org/projects/zookeeper/galaxy-zoo>; Milky Way Project, <https://www.zooniverse.org/projects/povich/milky-way-project>; and Chris J. Lintott, Kevin Schawinski, Anže Slosar, et al., "Galaxy Zoo: Morphologies Derived from Visual Inspection of Galaxies from the Sloan Digital Sky Survey," *Monthly Notices of the Royal Astronomical Society* 389 (3) (2008): 1179–1189, <https://doi.org/10.1111/j.1365-2966.2008.13689.x>.
- ⁵ Tin Kam Ho, "Random Decision Forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition* (IEEE, 1995), 278–282, <https://doi.org/10.1109/ICDAR.1995.598994>.
- ⁶ Duo Xu and Stella S. R. Offner, "Assessing the Performance of a Machine Learning Algorithm in Identifying Bubbles in Dust Emission," *The Astrophysical Journal* 851 (2) (2017), <https://doi.org/10.3847/1538-4357/aa9a42>.

- ⁷ Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, “Deep Learning,” *Nature* 521 (7553) (2015): 436–444; Colin M. Van Oort, Duo Xu, Stella S. R. Offner, and Robert A. Gutermuth, “CASI: A Convolutional Neural Network Approach for Shell Identification,” *The Astrophysical Journal* 880 (2) (2019): <https://doi.org/10.3847/1538-4357/ab275e>; and Duo Xu, Stella S. R. Offner, Robert Gutermuth, and Colin Van Oort, “Application of Convolutional Neural Networks to Identify Stellar Feedback Bubbles in CO Emission,” *The Astrophysical Journal* 890 (1) (2020), <https://doi.org/10.3847/1538-4357/ab6607>.
- ⁸ John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon, “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955,” *AI Magazine* 27 (4) (2006), <https://doi.org/10.1609/aimag.v27i4.1904>.
- ⁹ Sergios Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*, 2nd ed. (ScienceDirect, 2020).
- ¹⁰ Viviana Acquaviva, *Machine Learning for Physics and Astronomy* (Princeton University Press, 2023).
- ¹¹ David Donoho, “Data Science at the Singularity,” arXiv (2023), <https://doi.org/10.48550/arXiv.2310.00865>.
- ¹² Edward Kibblewhite, “Automated Measurement of Astronomical Photographs” (PhD diss., University of Cambridge, 1971).
- ¹³ W. Patrick McCray, “The Biggest Data of All: Making and Sharing a Digital Universe,” *Osiris* 32 (1) (2017), <https://doi.org/10.1086/693912>.
- ¹⁴ The Harvard Plate Stacks, “Our Collection,” <https://platestacks.cfa.harvard.edu/our-collection> (accessed December 16, 2025); and Harry Enke, Taavi Tuvikene, Detlef Groote, et al., “Archives of Photographic PLates for Astronomical USE (APPLAUSE): Digitisation of Astronomical Plates and Their Integration into the International Virtual Observatory,” *Astronomy & Astrophysics* 687 (2024), <https://doi.org/10.1051/0004-6361/202348793>.
- ¹⁵ McCray, “The Biggest Data of All,” citation therein.
- ¹⁶ Barbara A. Mikulski Archive for Space Telescopes, “HST Reports: Paper Counts,” <http://archive.stsci.edu/hst/bibliography/pubstat.html> (accessed December 16, 2025).
- ¹⁷ Alexander S. Szalay, “Jim Gray, Astronomer,” *Communications of the ACM* 51 (11) (2008): 58–65, <https://cacm.acm.org/research/jim-gray-astronomer>.
- ¹⁸ Tuan Dung Nguyen, Yuan-Sen Ting, Ioana Ciucă, et al., “AstroLLaMA: Towards Specialized Foundation Models in Astronomy,” arXiv (2023), <https://doi.org/10.48550/arXiv.2309.06126>.
- ¹⁹ Rui Pan, Tuan Dung Nguyen, Hardik Arora, et al., “AstroMLab 2: AstroLLaMA-2-70B Model and Benchmarking Specialised LLMs for Astronomy,” arXiv (2024), <https://doi.org/10.48550/arXiv.2409.19750>; and Kartheik G. Iyer, Mikael Yunus, Charles O’Neill, et al., “pathfinder: A Semantic Framework for Literature Review and Knowledge Discovery in Astronomy,” *The Astrophysical Journal Supplement* 275 (2) (2024), <https://doi.org/10.3847/1538-4365/ad7c43>.
- ²⁰ Sebastian Antony Joseph, Syed Murtaza Husain, Stella S. R. Offner, et al., “AstroVisBench: A Code Benchmark for Scientific Computing and Visualization in Astronomy,” arXiv (2025), <https://doi.org/10.48550/arXiv.2505.20538>.
- ²¹ Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, et al., “Hypothesis Generation with Large Language Models,” arXiv (2024), <https://arxiv.org/abs/2404.04326>.

- ²² Chenglei Si, Tatsunori Hashimoto, and Diyi Yang, “The Ideation-Execution Gap: Execution Outcomes of LLM-Generated versus Human Research Ideas,” arXiv (2025), <https://doi.org/10.48550/arXiv.2506.20803>.
- ²³ Ryan A. Loomis, Karin I. Öberg, Sean M. Andrews, et al., “Detecting Weak Spectral Lines in Interferometric Data through Matched Filtering,” *The Astronomical Journal* 155 (4) (2018), <https://doi.org/10.3847/1538-3881/aab604>.
- ²⁴ Jonah C. Rose, Paul Torrey, Francisco Villaescusa-Navarro, et al., “Introducing the DREAMS Project: DaRK mattEr and Astrophysics with Machine Learning and Simulations,” *The Astrophysical Journal* 982 (2) (2025), <https://doi.org/10.3847/1538-4357/adb8e5>; Philip F. Hopkins, Michael Y. Grudic, Kung-Yi Su, et al., “FORGE’d in FIRE: Resolving the End of Star Formation and Structure of AGN Accretion Disks from Cosmological Initial Conditions,” *The Open Journal of Astrophysics* 7 (2024), <https://doi.org/10.21105/astro.2309.13115>; and Dávid Guszejnov, Michael Y Grudić, Stella S. R. Offner, et al., “Effects of the Environment and Feedback Physics on the Initial Mass Function of Stars in the STARFORGE Simulations,” *Monthly Notices of the Royal Astronomical Society* 515 (4) (2022), <https://doi.org/10.1093/mnras/stac2060>.
- ²⁵ L. Prato and B. Gaensler, private communication, 2025.
- ²⁶ Guangping Li, Zujia Lu, Junzhi Wang, and Zhao Wang, “Machine Learning in Stellar Astronomy: Progress up to 2024,” arXiv (2025), <https://doi.org/10.48550/arXiv.2502.15300>; and Aldana Grichener, Mathieu Renzo, Wolfgang E. Kerzendorf, et al., “Nuclear Neural Networks: Emulating Late Burning Stages in Core-Collapse Supernova Progenitors,” *The Astrophysical Journal Supplement* 279 (2) (2025), <https://doi.org/10.3847/1538-4365/ade717>.
- ²⁷ Philip G. Breen, Christopher N. Foley, Tjarda Boekholt, and Simon Portegies Zwart, “Newton Versus the Machine: Solving the Chaotic Three-Body Problem Using Deep Neural Networks,” *Monthly Notices of the Royal Astronomical Society* 494 (2) (2020): 2465–2470, <https://doi.org/10.1093/mnras/staa713>.
- ²⁸ Adrian G. Adame, Janette Aguilar, Steven Ahlen, et al., “DESI 2024 VII: Cosmological Constraints from the Full-Shape Modeling of Clustering Measurements,” *Journal of Cosmology and Astroparticle Physics* 2025 (2025), <https://doi.org/10.1088/1475-7516/2025/07/028>.
- ²⁹ ChangHoon Hahn, Pablo Lemos, Liam Parker, et al., “Cosmological Constraints from Non-Gaussian and Nonlinear Galaxy Clustering Using the SimBIG Inference Framework,” *Nature Astronomy* 8 (2024): 1457–1467, <https://www.nature.com/articles/s41550-024-02344-2>.
- ³⁰ Patrick Mulvey and Jack Pold, “New Astronomy PhDs: What Comes Next,” American Institute of Physics, January 25, 2023, <https://www.aip.org/statistics/new-astronomy-phds-what-comes-next>.
- ³¹ Keith Poletti, Stella S. R. Offner, Rachel A. Ward, “Modeling Turbulent and Self-Gravitating Fluids with Fourier Neural Operators,” arXiv (2025), <https://doi.org/10.48550/arXiv.2507.23662>.
- ³² U.S. National Science Foundation, “National Artificial Intelligence Research Institutes,” <https://www.nsf.gov/funding/opportunities/national-artificial-intelligence-research-institutes> (accessed December 16, 2025).
- ³³ Andrew Ferguson, Marisa LaFleur, Lars Ruthotto, et al., “The Future of Artificial Intelligence and the Mathematical and Physical Sciences (AI+MPS),” arXiv (2025), <https://doi.org/10.48550/arXiv.2509.02661>.

³⁴ Ibid.

³⁵ Computer & Data Science Online, The University of Texas at Austin, “Graduate Certificate in Artificial Intelligence and Machine Learning,” <https://cdso.utexas.edu/caiml> (accessed December 15, 2025).

³⁶ Viktoria Cologna, Niels G. Mede, Sebastian Berger, et al., “Trust in Scientists and Their Role in Society across 68 Countries,” *Nature Human Behaviour* 9 (2025): 713–730, <https://doi.org/10.1038/s41562-024-02090-5>.

³⁷ Joyful E. Mdhuli, Charles Takalana, Ramasamy Venugopal, et al., “Astronomy as a Strategic Driver for Sustainable Development,” *Nature Astronomy* 9 (2025): 936–939, <https://www.nature.com/articles/s41550-025-02602-x>.

³⁸ Engin Akyol, Ahmed M. Hagag, Sobha Sivaprasad, and Andrew J. Lotery, “Adaptive Optics: Principles and Applications in Ophthalmology,” *Eye* 35 (1) (2020): 244–264, <https://doi.org/10.1038/s41433-020-01286-z>.

³⁹ NOIRLab, “The Charge-Coupled Device: Revolutionizing How Astronomers See the Universe after 50 Years of Development: What’s on the Horizon?” April 22, 2024, <https://noirlab.edu/public/blog/50-years-ccds>.

Language Is Not All You Need ... but Language, Probabilistic Programs & Bayesian Models of Cognition Will Get You Pretty Far

Joshua B. Tenenbaum

Since their origins in the 1950s, cognitive science and artificial intelligence have made slow but steady progress together toward a functional understanding of human intelligence. The arrival of large language models (LLMs) has upended this dynamic, with unprecedented commercial investment driven by the bet that superhuman AI could emerge simply from learning patterns in language at sufficient scale. While language is a singular tool for human thinking, there is far more to intelligence than language, as evidenced by how young children and nonhuman animals learn quickly and robustly even without language, and by the continuing jagged frontier of successes and failures in LLM-based AI. This essay considers another route to intelligent machines, grounded in principled theories and well-tested models of how minds and brains think before language and how learning language transforms thinking. By deploying AI breakthroughs in LLMs as models of language use – rather than as end-to-end models of intelligence – and connecting them to models of the thinking and learning our minds do prior to and independent of language, we have the opportunity not only to build more robust and efficient AI systems, but to rebuild the bridge between cognitive science and AI. This is a route to answering the biggest open questions about how human minds work, and with that understanding, making AI that positively impacts mental health, education, and society in ways unlikely to come from machine learning alone.

Language is the human superpower. Animals without language, including prelinguistic human children, can learn about the world, make reasonable inferences and decisions in scenarios they've never experienced, and even understand each other's actions and goals. But with language, these capacities vastly expand and deepen: We can learn from others, including people who lived long before we were born, and not only from our direct experience. We can cooperate and collaborate with others much more flexibly and effectively because we

can communicate our thinking with words. Language is also a medium of thought, for long-term planning and problem solving, and a substrate for associative memory and analogical retrieval. It lets us connect facts, stories, conversations, sights, sounds, tastes, and emotions – everything that we can talk about, however vaguely – to marshal from a lifetime of experience the knowledge and understanding we most need to be thinking with in any given moment.

The potential to capture these capabilities, and more, is what makes recent breakthroughs in AI language models so enticing. Artificial intelligence appears close to achieving the field’s oldest dream of an intelligent machine that you interact with as you would interact with another person – by talking to it. This is how AI has always presented itself to the broader world, dating back to Alan Turing’s 1950 paper on “Computing Machinery and Intelligence” – in which he famously proposed a chatbot-like imitation game (now known as the “Turing Test”) as a way to operationalize his vision of a thinking machine – or in science fiction films from 2001: *A Space Odyssey* (1968) to *Star Wars* (1977), *Star Trek* (2009), *Her* (2013), and many others.¹ Given their sci-fi-like capabilities, it’s no wonder that large language models (LLMs) are generating such excitement and enthusiasm in the general public, and that companies, governments, and investors of all kinds are putting significant capital and energy into them. We seem to be on the precipice of developing true thinking machines.

And yet, despite the many billions invested in LLM development over the last several years, and remarkable successes in building systems that can solve problems we used to think only humans could solve, there is something missing from this view of intelligence. LLM-based AI has been called “weird,” “alien,” “blurry,” and most tellingly “jagged”: that is, the frontiers of its capability do not advance smoothly in ways we have come to expect from how human intelligence develops; rather, we see world-class or even superhuman levels of AI expertise in some very challenging tasks and contexts, combined with surprisingly poor performance in much easier tasks.² When OpenAI released its first “reasoning-trained” model, o1, in December 2024, it scored as high as the top 0.1 percent of American high school students on the AIME, one of the hardest nationwide math contests, but it could not successfully play tic-tac-toe. A year later in December 2025, OpenAI’s latest GPT-5.2 model achieved a perfect (100 percent) score on the AIME but got less than 75 percent correct on a math contest designed for third and fourth graders.

As a cognitive scientist, I appreciate both why these systems seem so promising and why they fall short at achieving what we think of as human-like intelligence. The brain’s capacity for language is one of the most crucial, and most obvious, aspects of our species’ advanced cognition. But precisely because it’s so central to how we think, we tend to conflate language fluency with intelligence itself, leading to predictable biases against people with accents and people who stutter, and biases for people and tools we perceive as fluent.³ It’s no wonder, then, that we might tend to

think of LLMs, which can generate fluid prose, poetry, and code many times faster than humans, as exceptionally intelligent.

When you dig deeper from a cognitive science perspective, what you discover is more complicated. In a series of papers, my colleagues Evelina Fedorenko, Roger Levy, and their labs (in language and computational psycholinguistics) at MIT have shown that even small language models – such as OpenAI’s GPT-2, trained on more than a human lifetime’s worth of data but orders of magnitude less data than GPT-5 – can quantitatively predict language-specific neural responses in the human brain and behavioral signatures of sentence comprehension (such as the extra time it takes you to read a sentence with a surprising twist).⁴ These models capture “formal linguistic competence,” what you implicitly know about the rules of language that lets you interpret and produce sentences, but fall far short on tasks of “functional linguistic competence,” or any kind of thinking that requires synthesizing language facility with extralinguistic capabilities, like logical and causal reasoning, world modeling, and social cognition. These results, and other converging findings from LLMs trained on child-sized datasets and fMRI studies of how language in the brain is distinct from and interacts with other nonlinguistic brain systems, suggest that while language models are useful as models of human language processing, language alone is not all we need to build thinking machines or an accurate theory of how our minds work.⁵

I have spent the last twenty-five years building computational models of cognition, and specifically Bayesian models, which represent human thinking as a form of rational, approximate probabilistic inference, updating beliefs about the world by combining prior knowledge with newly observed evidence. So it will be no surprise that I believe developments in Bayesian cognitive science might hold the keys to building such a theory and a new generation of more human-like intelligence in AI. But I also believe that both cognitive science and artificial intelligence are at their best when they work in dialogue with each other.

These two fields grew up together in the 1950s and, for most of their history, made progress toward understanding intelligence by growing in conversation with each other. In recent years, this tradition of bidirectional and diverse exchanges has taken a back seat to Silicon Valley–driven scaling of a single architecture based on a single idea from machine learning: the transformer-based language model. While this will surely continue to produce economic and productivity gains, I believe that making further progress toward understanding the mind, and using that understanding to produce more rational and human-like forms of AI, is more likely to be achieved if we can rebuild the historical bridge between cognitive science and computer science.

As a step in this direction, my colleagues and I have been exploring how to think about the truly impressive advances in language modeling offered by LLMs through

a cognitive science lens. We are finding that language models can be integrated with both classic approaches to Bayesian cognitive modeling and recent developments in scalable Bayesian inference from probabilistic programming to open up a new and deeper understanding of human intelligence, and perhaps more intelligent AI.

Our starting point is the idea that intelligence starts with the capacity to model a changing world, to make inferences about the world's state and the likely effects of your actions in it, and to choose actions flexibly that you expect could achieve your goals. These capacities don't simply emerge over time as the result of learning to predict patterns in image sequences, speech streams, and forms of sensory data. Rather, they are core functions of brains that exist from the start and are the basis for how we can learn so much from so little information. This is true not just for humans but for animals whose brains are many times smaller than our own. Mice, for example, figure out how to move their bodies and manipulate objects in space to achieve their goals, despite their brains being about a thousand times smaller than ours and despite having no massive corpus of training data to draw on and no language capability. Like our brains, theirs are designed from the beginning to build internal models of the physical world that generalize to new situations beyond their direct experience, and to quickly update those models in response to their experiences. Mice can be tricked or trapped by humans thinking deeper and harder about physical mechanisms to catch them; but humans know very well that building a good mousetrap isn't easy. That's the point.

Human brains are, of course, capable of far more flexible thinking and rapid learning than mouse brains. This adaptive thinking begins well before children acquire language; it precedes and grounds language, and – together with our drive to understand and be understood by others – is partly what allows us to learn language so quickly and robustly. When children learn to read and write and begin to access all the knowledge that has accumulated culturally over generations, their intelligence and ability to learn accelerate in a reinforcing singularity that is the source of our species' linguistic superpower. But the seeds of these capabilities exist well before that. Intelligence, in other words, starts not with experience, data, or language, but with brains and minds built for rationality: constructing mental models to make good guesses and bets about the world and what one should do in it.

This is why I am skeptical that the “scaling thesis” for building AI, dominant in Silicon Valley today, is the best route to an understanding of the brain or to building machines with truly human-like intelligence. Unlike the human approach to *growing* intelligence from a seed of rationality, attempts to *scale* intelligence by scaling up data and compute start only with simple, “dumb-by-design” mechanisms of learning predictive patterns in data, and then hope everything else – world models, thinking and reasoning, social cognition, morality, and so on – emerges from it. We shouldn't be surprised when such models turn out to be jagged or brittle, require carefully curated training data to be effective across different scenarios,

generalize only weakly and unpredictably to settings outside of their training data (“out of distribution”), and become increasingly costly to update in response to new experiences.

These observations aren’t meant as a criticism of the scaling route – it’s a bet worth pursuing, at least if you have a trillion dollars. But the differences between the “growing up” and “scaling up” approaches might explain why LLMs can feel head-scratching even when they are working, and why it is worth also making bets on building systems that grow into intelligence the way a human being does.

This means starting with a computational architecture, as all brains do, for agency based on rational world modeling, and then introducing language in the way the human brain does, as a way of supercharging this architecture to make it vastly more flexible, generalizable, shareable, and scalable.

What precisely is an architecture for “rational world modeling”? This means constructing a mental model of our environment and – as a function of the state of the world we perceive, the actions we can take, and our goals – assessing the rewards of possible future states, the potential costs of actions, and the probabilities of our actions achieving those future states. From these components, we can compute the value of taking different actions by estimating which action sequences are likely to lead to high utility (rewards minus costs). We make these probabilistic inferences based on some combination of memories – how did similar actions in similar states turn out in the past? – and mental simulation, imagining based on our understanding of causal mechanisms how novel actions in novel states might unfold.

This architecture is not new and it is not uniquely human. It derives from the classical picture of rational agents as expected-utility maximizers, and in a deep sense it characterizes how all animals’ brains are designed to make good bets to achieve their goals.⁶ But in human minds, these processes are massively enriched and extended. Our world models include not just mental representations of the physical world but models of other agents in the world around us and, recursively, our approximations of their inner world models. Our utilities include not just basic survival needs, but our values, pleasures, and preferences, and those of other agents around us, including social norms and moral frameworks. And our decision and planning processes unfold not just in immediate fight-or-flight scenarios but over hours, days, and even years. When you add thinking and meta-reasoning to this architecture, it gets even more complex and powerful. We’re not just deploying basic world models anymore, we’re constructing and manipulating new models – including recursively and hierarchically composed models, and reduced models to help us plan efficiently – and extending our rational decision-making to our own internal computations, making good bets about how to use our computational resources.⁷

Language is partly what enables these distinctively human forms of world modeling, and what also transforms our thinking and learning in other ways. It gives our minds tools for marshaling, analogizing, recombining, and reinventing world

models from all the mental pieces we've ever built; tools for learning how to build new models, not only based on direct experience but, socially and culturally, from interactions with others and from accessing the vast stores of human knowledge accumulated over generations; and tools for imagining models of hypothetical or counterfactual ways the world could be or could have been but that have never appeared in any human being's experience.

How close are cognitive scientists to being able to implement this kind of architecture in a machine-executable model or a practical AI system? Not nearly as far along as the industry scaling route, but a lot closer than most of the world recognizes. The key pieces are in place and being prototyped on small scales, and we are ready to try larger-scale experiments that could be carried out with even a tiny fraction of the investment currently being poured into the scaling route.

This has become possible by building on three phases of research from the past twenty-five years of computational cognitive science. Starting in the late 1990s, under the banner of "Bayesian models of cognition," researchers began developing an approach to reverse-engineering the mind as an architecture for general-purpose and flexible rational world modeling. This work is reviewed in a 2024 book by Thomas Griffiths, Nick Chater, and me, written with many students, collaborators, and colleagues who have contributed to this research program and made it their own.⁸ Through their efforts, we now have a robust paradigm for building theoretically grounded, quantitatively predictive models of human learning and thinking that works in virtually every domain of cognition that behavioral scientists have studied, including causal reasoning, language acquisition and use, intuitive physics, decision-making and action understanding, social and moral cognition, and much more.

The most sophisticated of these Bayesian models have been built over the last decade, enabled by the advent of modern "probabilistic programming" tools starting around 2015. Probabilistic programs provide expressive languages for writing computational models of rational agents (natural or artificial) and their world models, as well as the inference algorithms agents use to perceive, reason, plan, learn, and communicate with their models.⁹ These expressive languages bring together the most powerful ideas from three influential paradigms for understanding intelligence as computation that have emerged over the history of AI and cognitive science: *symbolic* approaches, with their strengths of precision, abstraction, modularity, and composition; *neural-network* (or *differentiable-programming*) approaches, with their strengths of distributed representations, efficient parallelism, and gradient-based optimization; and *probabilistic* (or *Bayesian*) approaches with their strengths of rational inference and decision-making under uncertainty and risk. By allowing Bayesian modelers to take full advantage of symbolic and neural-network approaches and tools, probabilistic programming languages (PPLs) have made it possible to rapidly prototype, test, and scale cognitive models we could barely conceive of before 2010.

Probabilistic programming as a discipline is young but growing. In the last few years, we have seen these frameworks integrated with the same graphics processing unit (GPU) backends and transformer-based architectures that have driven the neural network revolution, and a proliferation of new PPLs are already being used to make Bayesian models of cognition more flexible, more efficient, or both.¹⁰ The arrival of code-trained LLMs and agentic AI coding environments (such as Anthropic's Claude Code) that can understand and integrate with PPLs will only accelerate this trend. Over the next two years, I expect we will increasingly see Bayesian models that used to take weeks to program manually, and were painfully slow to compute with, now being programmed in a day or less and tested almost instantly – much closer to the speed of thought.

More deeply, these tools are starting to make it possible to build the full capacities of natural language into our rational cognitive architectures. Though not standalone models of human intelligence, LLMs can be integrated with Bayesian cognitive models and probabilistic programs to build the first broad-coverage models of human thinking as it is extended and transformed by language. We can now test hypotheses about the many functional contributions language makes to cognition in ways that have not been possible before and, in the rest of this essay, I will briefly survey two ongoing efforts: one formalizes how language communicates structured thoughts in conversation, to build and use a world model jointly between two or more agents; the second formalizes how language can be used within a single mind (human or AI) to build flexible, bespoke world models for a new situation, by synthesizing, composing, and recombining pieces of other models marshaled from the agent's memory.

One long-standing view of the function of language is as an external medium for communicating human thoughts. Essentially, there is a mapping from an internal representation of thought into the external symbol systems that we call language, and learning language is learning this mapping. Everyone may have a different internal representational code, but presumably there is enough shared structure among minds that every child can learn an appropriate mapping. If this is right, we should be able to describe this process in a general way that starts with an underlying internal language designed for thinking – that is, making good guesses and bets about the world using the probabilistic models that are core features of the brain – and connecting this language of thought to externalizable symbol systems (spoken words or signs) via a probabilistic model of communication. This approach supports bidirectional mappings for both understanding the language of others and externalizing one's own thoughts.

In the paper “From Word Models to World Models,” by the cognitive science and AI researchers Lio Wong, Gabriel Grand, Alexander Lew, and others, we explore a method to do just that, which Wong calls “rational meaning construction.”¹¹ Our

work uses LLMs, but not as end-to-end thinking systems. Instead, they are used to model this superpower that language gives human minds: a tool for mapping between the external language of a conversation with other minds and the mind's own internal languages of thought.

More technically, Wong and colleagues show how to integrate LLMs with Bayesian models written in a PPL to give an architecture for “language-informed thinking,” a proposal that contrasts dramatically with today's AI industry scaling route. Rather than attempting to learn all of thinking from patterns in language, the focus is an agent architecture that starts with a substrate designed for rational thinking: a probabilistic language of thought (PLOT), which we formally instantiate using probabilistic programs that express an AI agent's foundational knowledge (or priors) over possible worlds. Then we use LLMs to translate natural language into these probabilistic programs, which in turn are used to perform context-sensitive Bayesian inference over distributions of possible worlds to infer what is likely to be true or likely to be a good answer to any question or decision, updated from an agent's priors in light of what they have observed or been told in a conversation.¹² Relative to today's frontier AI models, we can use much smaller, more data- and energy-efficient LLMs to implement this mapping – perhaps closer to the resources of the human brain – because the LLM is only being asked to understand language, not to implement (or imitate) reasoning.

I don't have space here to cover all the recent developments that demonstrate the promise of this approach, but I will highlight just one. In a study published in January 2025, psychology and machine-learning researcher Luca Schulze Buschhoff and colleagues evaluated how well state-of-the-art multimodal (vision-trained) LLMs captured the judgments of human participants in simple commonsense settings: intuitive physics, causal reasoning, and intuitive psychology.¹³ The models struggled to match human patterns of reasoning, such as predicting whether a stack of blocks shown in an image is stable, even though they were able to perform basic visual operations on the same scenes, such as recognizing the colors of the blocks.

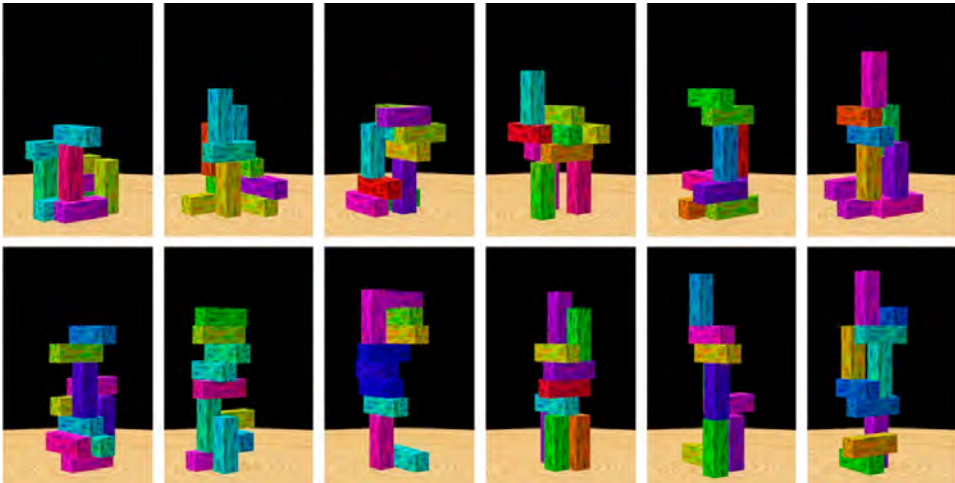
These results are telling because cognitive scientists have long known how to capture human judgments in these tasks using Bayesian models written as probabilistic programs, and indeed the tasks studied were designed originally to test these models. For example, in intuitive physics, Peter Battaglia and Jessica Hamrick built a Bayesian model for how people make inferences about similar block-stacking scenarios like those depicted in Figure 1.

Based on reconstructing 3D scenes from images and simulating their next few moments in a video-game-style approximate physics engine, their model captured people's judgments about the stability of towers and many other scenarios like these, with high quantitative accuracy that has yet to be matched by AI systems.¹⁴

It is worth emphasizing how much this approach differs from traditional machine learning. The Bayesian model requires no task-specific training with

Figure 1

Example Block-Stacking Scenarios Designed to Study Human Intuitive Judgments of Physical Stability



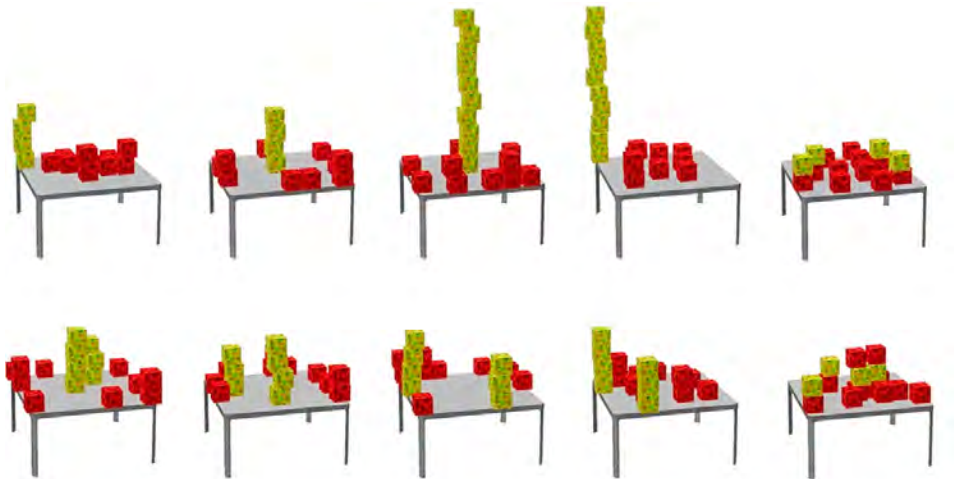
Source : Peter W. Battaglia, Jessica B. Hamrick, and Joshua B. Tenenbaum, “Simulation as an Engine of Physical Scene Understanding,” *Proceedings of the National Academy of Sciences* 110 (45) (2013): 18327–18332.

external feedback signals to judge stability. It makes predictions purely from its own internal simulations – a form of “imagination.” This allows it to generalize immediately, as people can, to tasks very different from those encountered before. For example, given tabletop scenes of red and yellow blocks like those in Figure 2, we can ask people to judge how likely it is that bumping the table will lead to more red blocks or yellow blocks falling onto the floor. The same probabilistic simulation approach captures these judgments almost as well as it does the much more familiar task of judging stability.

This is also where we can see the power of bringing language into the picture, as well as the idea of rational meaning construction at work. What if instead of being asked to make an inference based on the images above, a human or AI agent was given only a verbal description of a scene? If I tell you that there are several tall stacks of red blocks near the center of a table, and a few short stacks at the table’s edge, and less than half of the blocks in the short stacks are yellow, you would probably predict that a bump to the table is likely to cause more red blocks to fall off. If instead I told you that only one of the tall stacks had red blocks, you might instead predict more yellow blocks would fall off. You make these inferences not

Figure 2

Example Scenarios Designed to Test Probabilistic Simulation Models of Novel Intuitive Physics Judgments



Source : Illustration, produced by Peter Battaglia, courtesy of the author.

by inspecting the actual visual scene in front of you but by constructing an imagined scene in your mind’s eye.

AI and cognitive science researchers Cedegao Zhang, Lio Wong, Gabriel Grand, and I captured this kind of language-based thinking in our first empirical demonstration of the rational meaning construction approach. We used an LLM to translate text descriptions of red and yellow block scenes into code for probabilistic simulation models, which can then be run in the same simulation-based inference mode as before.¹⁵ Our experiments included both short, vague descriptions such as “a tall stack of red blocks,” as well as more complex ones, with multiple stacks of blocks or an exact number stacked in more precise locations. By averaging the results across a few of these translations and a few simulations from each synthesized probabilistic model, we could quantitatively predict people’s judgments almost as well as in the original, image-based study.

Recall that the Schulze Buschoff study showed that cutting-edge vision-language transformers performed far worse at much simpler tasks.¹⁶ And yet, when paired with domain-specific Bayesian world models, even small language models were able to translate complex descriptions of scenes into reasonably accurate code simulation models that make similar judgments to people about the same scenarios. To

me this shows both the power of language to let us think flexibly about new situations, and the value of not identifying language with all of thinking.

This work is still very preliminary, and intuitive physics is just one domain of cognition. But we are already seeing similar advantages for rational meaning construction in other settings of complex language-informed thinking: for instance, in modeling how people understand agents' beliefs, desires, and preferences from observing their interactions and communication, or how they form new causal world models in gameplay from combinations of experience and advice from previous players.¹⁷ At this point, I'd say it's a reasonable bet that somehow combining Bayesian inference in probabilistic programs as the rational core of human thinking with code-trained language models to capture the way people spin words into these world models can take us far in both computational cognitive science and AI.

And yet if we are looking for a full theory of human cognitive architecture, any such approach still faces major challenges of scaling and generality. Most immediately, in each domain where we have built highly predictive quantitative models of human reasoning, the probabilistic code for world modeling has been written at least partly by human cognitive scientists. But in the real world, no external designer is there to program our mind's world models for us. How would all this code (or its analog) get into people's heads? And could we engineer an AI system that generates these capabilities for itself without needing to be programmed even partially from the outside?

Put another way, could we build cognitive architectures with the same inputs and outputs as today's LLMs – which, like human minds, can extend generally and flexibly to any scenario describable in words – but with internal processes that truly mirror the mind's operations? Such a system would have to be able to construct world models on its own, at least in all the situations and at the level of depth and explicitness that humans can imagine, and then reason coherently, rationally, and efficiently at least as well as humans do.

Confronting this challenge is the goal of the second line of work I want to talk about briefly. An effort led by cognitive scientists Tyler Brooke-Wilson, Lio Wong, and Katie Collins has begun exploring the hypothesis that the mind operates with what we call a “model synthesis architecture” (MSA). Brooke-Wilson originally proposed and motivated the MSA hypothesis in his doctoral dissertation at MIT, and this team (which I am part of) is now trying to implement it computationally and test it behaviorally.¹⁸ We posit that human minds, rather than walking around with a giant monolithic probabilistic codebase expressing everything they know about the world, construct small bespoke world models on the fly, as needed, incorporating only the variables and causal dependencies that might be relevant to a given situation and the inference problem they are currently trying to solve. These ad hoc, situation-specific world models take the form of probabilistic programs that

support efficient probabilistic inference, just as in all the examples described earlier. But this code is now synthesized online by the agent via a process of “marshalling,” or assembling and adapting fragments of previously constructed models based on similar problems and situations recalled from memory. This architecture thus allows minds that are computationally limited (in processing, memory, time, and other resources) to reason coherently and rationally in local contexts while drawing on globally relevant considerations from across their experiences, knowledge, and beliefs.

We have implemented a simple first prototype of an MSA using a combination of LLMs and PPLs.¹⁹ As in rational meaning construction, where we used LLMs to capture how people use language as a tool for *externalizing* and communicating structured thoughts, here we use code-trained LLMs to capture how minds use language in a different way – as an *internal* tool for general-purpose associative memory, analogical retrieval, and mapping. These associations and analogies span multiple forms of mental representation: general, often vague, background knowledge; informal “word models” of specific situations; and more formal symbolic (PPL-based) world models. These are just the capacities needed in a system for bespoke model construction that supports broad coverage yet locally coherent rational inference.

Our MSA prototype is being tested in a classic domain of probabilistic causal reasoning: making inferences about the outcomes and traits of players in team sports matches, such as tug-of-war, doubles ping-pong, or canoe racing.²⁰ For instance, you might be told the outcomes of several races between teams of participants (“In the first race, Val and Gale lost to Sam and Kay. Val and Harper beat Sam and Ness in the second race . . .”) and be asked to judge how strong you think Sam is, or how hard they tried in the first race. People can make graded probabilistic judgments like these systematically and reliably – and the MSA can explain how they do so.

Reasoning in this architecture proceeds in several stages, moving from a natural-language problem description to PPL code generation to probabilistic inference in the generated world model. Each stage is implemented through a combination of language model proposals and probabilistic reweighting of sampled model traces. First, the problem description is parsed into fragments of code expressing the observed facts and the questions to be answered. The observation “In the first race, Val and Gale lost to Sam and Kay” might be parsed into the expression

```
condition(lost({team1: ['val', 'gale'], team2: ['sam', 'kay'], race: 1}))
```

and the question “How hard did Sam try in the first race?” into

```
query: effort_level_in_race({athlete: 'sam', race: 1}).
```

This step is effectively the same as in rational meaning construction, and similar to some linguistic views of “understanding” meaning in language as a mapping into a formal predicate-based representation, but with mappings that are more pragmatic and situation-specific. The predicates have names in English, with all

the associations those carry, but they are not yet meaningfully grounded in any world model. They become formally meaningful only in the context of a bespoke probabilistic program the architecture will synthesize for reasoning about this problem. That synthesis proceeds next, by marshaling relevant general knowledge from memory and assembling it into informal (natural-language) descriptions of hypothetical dependencies between variables, along with informal probabilities. For example, the system might posit that *“Intrinsic strength varies widely from athlete to athlete, but let’s imagine that players tend to be weak, average or strong, and uniformly around that . . . ,”* or *“The amount of effort that an athlete puts into any given race is a continuous parameter. . . . This effort is somewhat contingent on an athlete’s intrinsic strength, because stronger athletes probably tend to be more likely to put in extra effort.”* Finally, these informal model fragments are used to synthesize a formal code-based world model in a PPL; this grounds the predicates expressing the questions posed and supports coherent, rational inferences to answer them.

While still preliminary, this approach does appear capable of making judgments quantitatively similar to humans on these tasks – and outperforming simple language models on their own. In particular, the MSA prototype appears to better capture how people reason coherently in novel situations, when suddenly called upon to consider a variable or factor that just wasn’t in their mental model before. We asked a second set of participants to propose new explanatory variables or factors that might affect match outcomes or how one would judge the strength of a player. Some human responses we got included: *“Avery seems to have come down with a stomach virus between rounds, but has decided to compete anyway.”* *“Kay didn’t get enough sleep last night and can barely stay awake during the race.”* *“Ness took an energy drink and it just kicked in to give them extra energy for their matches.”*²¹ We then gave these same facts back to the first set of participants, in addition to match outcomes, and asked them to judge players’ strengths conditioned on this new information. The MSA handles these scenarios too, following the same process as above. It first proposes, informally in natural language, causal mechanisms that link the new information to what is already known, and then translates those proposals into formal PPL representations of how its world model must be expanded. Probabilistic inference in the expanded model yields results that, although messy like the real world, are much closer to human judgments than reasoning in natural language (the pure LLM) alone.

This is just a first demonstration, but it points to the larger promise of the MSA approach. Because the models synthesized in this architecture are small ones – instantiating just the relevant variables needed for a given problem – inference and decision-making are tractable and efficient in a way they could never be if we had one big symbolic model for the whole world in our heads. Because the models are constructed on the fly, we can in principle reason about any situation, so long as we can retrieve at least some relevant background information that can connect the question to the observations we’ve made.

The ability to assemble the relevant variables from background experience and check them against the specific, present context also illustrates more broadly the functions natural language plays in human thinking beyond merely communicating ideas. On the one hand, language provides the substrate for almost everything we come to know beyond our direct experience – everything learned explicitly or implicitly, abstractly or concretely, from people telling us things, or reading, or other cultural means. At the same time, language is the basis for a general content-addressable associative memory that indexes this world knowledge together with every memory of the symbolic worlds or fragments of models we’ve built before, and allows us to efficiently retrieve candidate variables, relations between variables, and symbolic model components useful for the current task at hand.

Implementing all these functions in the weights of code-trained LLMs, as our prototype does, is not the only way to build an MSA. But it does mirror something of how neuroscientists think knowledge is represented in the weights of the brain’s long-term memory: not always or even usually in an explicit and structured form, but in a distributed and implicit manner that can be marshaled into structured world models to support coherent rational reasoning as and when needed for a task.

In sum, this architecture may offer an approach to building a serious and scalable machine-executable theory of how human language and rational thought interact, consistent with our understanding of their evolutionary, developmental, and neural bases. The MSA uses the foundation of probabilistic programming to capture how people reason coherently about the world, whether in long-standing core domains such as intuitive physics and intuitive psychology or flexibly beyond those to all the situations we can think about in small bespoke models. And it incorporates breakthroughs in language modeling from artificial intelligence, not as an end-to-end paradigm for all of intelligence, but as a tool for all the ways that language extends, transforms, and scales the human capacity for rationality, mental model building, and coherent reasoning.

I want to close this essay by looking forward to the biggest open scientific questions at the intersection of human and artificial intelligence and the societal impacts that may come from answering them. There are many open questions about the model synthesis architecture, chief among them are whether and in what sense it actually resembles how the brain works. Even if it is right in broad strokes, the actual mechanics of this architecture will surely be more complex – in how and when the brain synthesizes explicit world models across different domains of cognition, what form its languages of thought takes and how they interact with natural language, and how model synthesis and inference build on and interleave in online thinking.

But to the extent that these ideas do capture something fundamental about how the brain works – how intelligence grows from a foundation of rationality, and how language both emerges from and expands these core capabilities – they could have

significant and wide-ranging implications for fields from computing and psychology to education and economics.

If we can build AI grounded in a functional understanding of human intelligence, we may develop far more efficient and effective systems, able to reason coherently and act rationally using much less data, computation, and energy. By understanding how the brain works as an organ of rationality, and how that function can break down, we may develop therapies to address mental illness and improve mental health by targeting specific broken cerebral processes. By better understanding how our minds grow, we may better diagnose and treat developmental disabilities. And by better understanding how people communicate and learn, and interpret information at scale, we may develop far more predictive economic models and effective educational techniques. In the process of making these discoveries, we may even answer some of the biggest questions our field has long puzzled over: What kind of computation is cognition? How do these computations work mechanistically in the brain? And how did our minds arise through evolution, culture, and development?

In my view, these questions and many important ones like them are unlikely to be answered purely by continuing along the LLM scaling route or by publishing bespoke cognitive models alone. To a cognitive scientist, the way large language models are used today in AI, where language is the source of thinking and everything else is a tool that these models use to become more rational, is backward. Instead, we should see language as a tool – a superpower, but still a tool – that emerges out of already rational minds and enhances and extends them. Likewise, we need to move past thinking of symbols in terms of the brittle, intractable, hard-coded formalisms of classical AI (or earlier Bayesian cognitive science). Symbolic systems need not be fixed, massive, monolithic codebases. Instead, they can be small, bespoke, and transient models, fast to build, quick to do inference in, and cheap enough to cast aside. Code doesn't need to be "written by hand" but can be constructed evolutionarily: learned from experience (including language), and rewritten, reused, and recomposed by the mind itself.

By combining the tools of modern language models, probabilistic programming, and Bayesian cognitive science, we have been able to build prototypes of this way of thinking, much in the same way we view the mind as marshaling intelligence out of multiple different systems that build on our core rationality. We are only at the beginning, but I hope this work could be the basis for rebuilding the bridge between the fields of artificial intelligence and cognitive science and their working together toward a foundational architecture that achieves human-like language and thought. That may just lead to AI systems that really are the thought partners we've always wanted, that make us truly smarter and better off, and that start to answer some of the fundamental questions about how our minds work and where our intelligence comes from.

AUTHOR'S NOTE

This essay is based on a keynote talk, “Scaling Intelligence the Human Way: Building a Shared Future for AI and Cog Sci,” I presented on August 1, 2025, at the 47th Annual Meeting of the Cognitive Science Society. It represents joint work with many collaborators, but especially Lio Wong and Tyler Brooke-Wilson, who originated the ideas of rational meaning construction and model synthesis architectures (respectively) in their PhD work and who continue to lead the development of this research program. Katherine Collins provided essential contributions as a colead on much of this work, and key contributions also came from Alexander Lew, Gabriel Grand, Jacob Andreas, Vikash Mansinghka, Cedegao Zhang, Lance Ying, Tan Zhi-Xuan, Kaya Stechly, Tobi Gerstenberg, Noah Goodman, and Timothy O'Donnell. I am grateful for the support of an AI2050 Senior Fellowship from Schmidt Sciences, a MacArthur Fellowship, and grants from the Office of Naval Research, the Air Force Office of Scientific Research, the U.S. National Science Foundation, and the MIT Siegel Family Quest for Intelligence. Thanks also to Gabe Stein and Jennifer Sunoo for writing assistance, and to Mira Bernstein, Laura Schulz, and James Manyika for comments on this essay.

ABOUT THE AUTHOR

Joshua B. Tenenbaum, a Member of the American Academy since 2020, is Professor of Computational Cognitive Science in the Department of Brain and Cognitive Sciences (BCS) at MIT, a Principal Investigator at MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL), and the Director of Science at the MIT Siegel Family Quest for Intelligence. He was named a MacArthur Fellow in 2019 in recognition of his contributions to mathematical psychology and Bayesian cognitive science, and is the recipient of the Troland Award from the National Academy of Sciences and the Howard Crosby Warren Medal from the Society of Experimental Psychologists.

ENDNOTES

- ¹ Alan M. Turing, “Computing Machinery and Intelligence,” *Mind* 49 (236) (1950): 433–460, <https://doi.org/10.1093/mind/LIX.236.433>.
- ² Andrej Karpathy (@karpathy), “Jagged Intelligence The word I came up with to describe the (strange, unintuitive) fact that state of the art LLMs can both perform extremely impressive tasks (e.g. solve complex math problems) while simultaneously struggle with some very dumb problems. E.g. example from two days ago - which number is bigger, 9.11 or 9.9? Wrong. <https://t.co/3C7pCdBSHQ>,” X (formerly Twitter), July 25, 2024, 1:50 p.m., <https://x.com/karpathy/status/1816531576228053133>.
- ³ Kyle Mahowald and Anna A. Ivanova, “Google’s Powerful AI Spotlights a Human Cognitive Glitch: Mistaking Fluent Speech for Fluent Thought,” *The Conversation*, June 24, 2022, <https://doi.org/10.64628/AAI.wvs9mp5dy>.
- ⁴ Martin Schrimpf, Idan Asher Blank, Greta Tuckute, et al., “The Neural Architecture of Language: Integrative Modeling Converges on Predictive Processing,” *Proceedings of the National Academy of Sciences* 118 (45) (2021): e2105646118, <https://doi.org/10.1073>

- /pnas.2105646118; Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, et al., “Dissociating Language and Thought in Large Language Models,” *Trends in Cognitive Sciences* 28 (6) (2024): 517–540, <https://doi.org/10.1016/j.tics.2024.01.011>; and Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, et al., “On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior,” arXiv (2020), <https://doi.org/10.48550/arXiv.2006.01912>.
- ⁵ Alex Warstadt, Aaron Mueller, Leshem Choshen, et al., “Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora,” *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, ed. Alex Warstadt, Aaron Mueller, Leshem Choshen, et al. (Association for Computational Linguistics, 2023), 1–6, <https://doi.org/10.18653/v1/2023.conll-babylm.1>; and Colton Casto, Anna Ivanova, Evelina Fedorenko, and Nancy Kanwisher, “What Does It Mean to Understand Language?” arXiv (2025), <https://doi.org/10.48550/arXiv.2511.19757>.
- ⁶ Michael Tomasello, *The Evolution of Agency: Behavioral Organization from Lizards to Humans* (MIT Press, 2022).
- ⁷ Samuel J. Gershman, Eric J. Horvitz, and Joshua B. Tenenbaum, “Computational Rationality: A Converging Paradigm for Intelligence in Brains, Minds, and Machines,” *Science* 349 (6245) (2015): 273–278, <https://doi.org/10.1126/science.aac6076>; and Falk Lieder and Thomas L. Griffiths, “Resource-Rational Analysis: Understanding Human Cognition as the Optimal Use of Limited Computational Resources,” *Behavioral and Brain Sciences* 43 (2019), <https://doi.org/10.1017/S0140525X1900061X>.
- ⁸ Thomas L. Griffiths, Nick Chater, and Joshua B. Tenenbaum, *Bayesian Models of Cognition: Reverse Engineering the Mind* (MIT Press, 2024), <https://mitpress.mit.edu/9780262049412/bayesian-models-of-cognition>. This work in turn builds on pioneering rational analyses of generalization, categorization, and memory by cognitive scientist Roger Shepard and psychologist John Anderson in the late 1980s and early 1990s, and a computational framework for the study of vision pioneered by neuroscientist David Marr in his book *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (W. H. Freeman and Company, 1982).
- ⁹ For an introduction to probabilistic programming in cognitive science and artificial intelligence, see Noah D. Goodman, Joshua B. Tenenbaum, and the ProbMods Contributors, *Probabilistic Models of Cognition*, 2nd ed. (2016), <https://www.probmods.org>; and The Gen.jl Team, “Gen Tutorials,” <https://www.gen.dev/tutorials> (accessed January 15, 2026).
- ¹⁰ Eli Bingham, Jonathan P. Chen, Martin Jankowiak, et al., “Pyro: Deep Universal Probabilistic Programming,” *The Journal of Machine Learning Research* 20 (28) (2019): 1–6, <https://jmlr.org/papers/v20/18-403.html>; Marco F. Cusumano-Towner, Feras A. Saad, Alexander K. Lew, and Vikash K. Mansinghka, “Gen: A General-Purpose Probabilistic Programming System with Programmable Inference,” in *PLDI 2019: Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation* (Association for Computing Machinery, 2019), 221–236, <https://doi.org/10.1145/3314221.3314642>; Kartik Chandra, Tony Chen, Joshua B. Tenenbaum, and Jonathan Ragan-Kelley, “A Domain-Specific Probabilistic Programming Language for Reasoning About Reasoning (or: A Memo on Memo),” *Proceedings of the ACM on Programming Languages* 9 (OOPSLA2) (2025): 784–814, <https://doi.org/10.1145/3763078>; and McCoy R. Becker, Alexander K. Lew, Xiaoyan Wang, et al., “Probabilistic Programming with Programmable Variational Inference,” *Proceedings of the ACM on Programming Languages* 8 (PLDI) (2024): 2123–2147, <https://doi.org/10.1145/3656463>.

- ¹¹ Lionel Wong, Gabriel Grand, Alexander K. Lew, et al., “From Word Models to World Models: Translating from Natural Language to the Probabilistic Language of Thought,” arXiv (2023), <https://doi.org/10.48550/arXiv.2306.12672>.
- ¹² Technically this translation process itself is a probabilistic inference. The meaning function parameterized by a large language model encodes a distribution on likely meanings of a sentence in the form of a distribution over expressions in the probabilistic language of thought. See *ibid.*, note 14.
- ¹³ Luca M. Schulze Buschoff, Elif Akata, Matthias Bethge, and Eric Schulz, “Visual Cognition in Multimodal Large Language Models,” *Nature Machine Intelligence* 7 (1) (2025): 96–106, <https://doi.org/10.1038/s42256-024-00963-y>.
- ¹⁴ Peter W. Battaglia, Jessica B. Hamrick, and Joshua B. Tenenbaum, “Simulation as an Engine of Physical Scene Understanding,” *Proceedings of the National Academy of Sciences* 110 (45) (2013): 18327–18332. See also Kevin A. Smith, Jessica B. Hamrick, Adam N. Sanborn, et al., “Intuitive Physics as Probabilistic Inference,” in *Bayesian Models of Cognition*, ed. Griffiths, Chater, and Tenenbaum, ref. 10. It might seem strange to posit that the brain contains a physics simulation engine, but Nancy Kanwisher’s lab at MIT has used functional neuroimaging to show that indeed it might, in a candidate “physics network” spanning frontal and parietal regions that intriguingly are also involved in action planning and perhaps how we understand others’ actions and even novel tool use. Importantly, neither people’s judgments nor those of the Bayesian simulation-based models are completely accurate. Because they only approximately and probabilistically infer the underlying physical scene, they show systematic “physical illusions,” such as seeing stacks as unstable that in fact are not. The model and people make these errors in the same ways, seeing the same physical illusions, providing further evidence that this is how intuitive physics works (and doesn’t always work) in our brains.
- ¹⁵ Cedegao Zhang, Lionel Wong, Gabriel Grand, and Josh Tenenbaum, “Grounded Physical Language Understanding with Probabilistic Programs and Simulated Worlds,” *Proceedings of the Annual Meeting of the Cognitive Science Society* 45 (2023), <https://escholarship.org/uc/item/7018f2ss>.
- ¹⁶ Schulze Buschoff, Akata, Bethge, and Schulz, “Visual Cognition in Multimodal Large Language Models.”
- ¹⁷ Chris L. Baker, Rebecca Saxe, and Joshua B. Tenenbaum, “Action Understanding as Inverse Planning,” *Cognition* 113 (3) (2009): 329–349, <https://doi.org/10.1016/j.cognition.2009.07.005>; Julian Jara-Ettinger, Hyowon Gweon, Laura E. Schulz, and Joshua B. Tenenbaum, “The Naïve Utility Calculus: Computational Principles Underlying Commonsense Psychology,” *Trends in Cognitive Sciences* 20 (8) (2016): 589–604, <https://doi.org/10.1016/j.tics.2016.05.011>; Tan Zhi-Xuan, Jordyn Mann, Tom Silver, et al., “Online Bayesian Goal Inference for Boundedly Rational Planning Agents,” in *Advances in Neural Information Processing Systems* 33, ed. Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, et al. (Curran Associates, Inc., 2020), 19238–19250, https://proceedings.neurips.cc/paper_files/paper/2020/hash/df3aebc649f9e3b674eeb790a4da224e-Abstract.html; Lance Ying, Katherine M. Collins, Megan Wei, et al., “The Neuro-Symbolic Inverse Planning Engine (NIPE): Modeling Probabilistic Social Inferences from Linguistic Inputs,” arXiv (2023), <https://doi.org/10.48550/arXiv.2306.14325>; Lance Ying, Ryan Truong, Katherine M. Collins, et al., “Language-Informed Synthesis of Rational Agent Models for Grounded Theory-of-Mind Reasoning On-The-Fly,” arXiv (2025), <https://doi.org/10.48550/arXiv.2506.16755>; Pedro A. Tsividis, Joao Loula, Jake Burga, et al., “Human-Level Reinforcement Learning through Theory-Based Modeling, Exploration, and Planning,” arXiv (2021),

<https://doi.org/10.48550/arXiv.2107.12544>; and Cédric Colas, Tracey Mills, Ben Prystawski, et al., “Language and Experience: A Computational Model of Social Learning in Complex Tasks,” arXiv (2025), <https://doi.org/10.48550/arXiv.2509.00074>.

- ¹⁸ Tyler Brooke-Wilson, “Bounded Rationality as a Strategy for Cognitive Science” (PhD diss., Massachusetts Institute of Technology, 2023).
- ¹⁹ Lionel Wong, Katherine M. Collins, Lance Ying, et al., “Modeling Open-World Cognition as On-Demand Synthesis of Probabilistic Models,” arXiv (2025), <https://doi.org/10.48550/arXiv.2507.12547>.
- ²⁰ Based on the Bayesian Tug of War from Noah D. Goodman, Joshua B. Tenenbaum, and Tobias Gerstenberg, “22: Concepts in a Probabilistic Language of Thought,” in *The Conceptual Mind: New Directions in the Study of Concepts*, ed. Eric Margolis and Stephen Laurence (MIT Press, 2015), 623–654, <https://doi.org/10.7551/mitpress/9383.003.0035>.
- ²¹ Wong, Collins, Ying, et al., “Modeling Open-World Cognition as On-Demand Synthesis of Probabilistic Models.”

Knowledge-Centric AI for Scientific Discovery

Carla P. Gomes

Recent breakthroughs in artificial intelligence have been driven primarily by data-centric approaches, including deep learning and large language models. However, despite their remarkable success, relying solely on purely data-driven methods has intrinsic limitations for scientific discovery. In this essay, I argue for knowledge-centric AI, reasoning grounded in scientific principles, domain structure, and constraints while learning from data. In our research at Cornell University and through interdisciplinary collaborations with researchers across institutions, this agenda has been shaped by data-limited scientific challenges, which has led us to develop knowledge-centric approaches, such as deep reasoning networks, which build interpretable, domain-aligned latent spaces and enforce differentiable constraints end-to-end. Knowledge-centric AI has enabled scientific advances, from automated crystal-structure phase mapping that uncovered high-performing alloyed mixtures to joint species–distribution models that inform conservation, while also delivering general AI methodologies. By aligning with the scientific method and integrating reasoning, learning, and principled experimentation, knowledge-centric AI can catalyze scientific innovation. In our work, this perspective is grounded in computational challenges central to a sustainable future.

Through breakthroughs such as AlphaGo, AlphaFold, and ChatGPT, artificial intelligence is at the cusp of transforming our lives.¹ These are exciting times, as the field moves closer to artificial general intelligence, replicating general purpose human cognition.² At the same time, humanity faces significant global challenges, including climate change and biodiversity loss, which threaten ecosystems and human well-being worldwide. Recognizing these issues, the United Nations' 2030 Agenda for Sustainable Development formulated seventeen Sustainable Development Goals to end poverty, protect the planet, and ensure prosperity for all.³ Artificial intelligence holds great promise for advancing such goals, but realizing its positive impact requires deliberate effort and careful stewardship. Since the late 2000s, we have helped establish and nurture the field of computational sustainability with the belief that computational scientists can and should tackle societal and environmental problems while simultaneously driving founda-

tional advances in computer science.⁴ Our focus is on developing and advancing AI methodologies to accelerate scientific discovery to help address sustainability challenges.

Recent AI breakthroughs have been driven by data-centric approaches, notably deep learning and large language models, but relying on these alone has intrinsic limitations for scientific discovery.⁵ Here, I argue that knowledge-centric AI is key to advancing scientific discovery, as it addresses two unique challenges for scientific discovery and exploration.

First, high-quality experimental data are often time-consuming to obtain and interpret, leading to relatively data-limited settings. Second, one needs to leverage existing scientific laws and principles, which allow reaching far beyond any specific training distribution. For example, Newton's laws of motion and gravity were based on experiments and observations conducted on Earth. The leap to apply the same laws and principles to celestial objects is a beautiful example of the ability of scientific laws to generalize across vastly different contexts. Scientists are constantly pursuing such general principles and insights. They use existing scientific theories to guide their exploration – looking for small inconsistencies with experimental data to suggest modifications and refinements – to enable broader applicability of the theories. For instance, Einstein's work on relativity adapts Newton's laws to settings in which speeds approach the speed of light and extreme gravitational forces affect Euclidean notions of time and space. Similarly, AI-driven scientific discovery needs to leverage existing scientific laws and theories to uncover new phenomena and develop new theories. This essay emphasizes the power of knowledge-centric and first-principles reasoning: by integrating reasoning, learning, and experimentation, artificial intelligence can become a catalyst for discovery.

Science represents one of the purest expressions of human intellect. Psychologist Daniel Kahneman characterizes human thinking as comprising two systems.⁶ System 1 is fast, automatic, and pattern driven, excelling at perception and intuitive recognition. In contrast, System 2 is slow, deliberate, and responsible for logical reasoning and complex decision-making. Over the past decade, AI has made spectacular progress on System 1–like capabilities through data-driven approaches. However, scientific discovery leans heavily on System 2, knowledge-driven reasoning, wherein limited data are enriched with first-principles reasoning about prior scientific knowledge.⁷

Scientific AI agents (SciAIs) – artificial systems that can devise, conduct, and analyze scientific research with some degree of autonomy – should therefore combine data-driven learning with explicit knowledge-driven reasoning. The line between data and knowledge can be subtle, but it has critical consequences. Crucially, the belief that “enough data” can fully replace knowledge is misguided. The conceptual world is combinatorial and complex, and many scientific regimes are data limited

but knowledge rich. To produce reliable, novel insights, SciAIs need representations that enable explanation, verification, and generalization.

SciAIs should emulate the scientific process, blending multimodal learning with principled reasoning, choosing representations that scale, and manipulating knowledge effectively. They must predict and discover well beyond their training distribution, interpret results, and identify causal relationships. Achieving true generalization and creativity in science requires scientific common sense and deductive, counterfactual, and abductive reasoning, supported by domain-specific inductive biases and hybrid data-knowledge methods that adapt to scarce or atypical data.

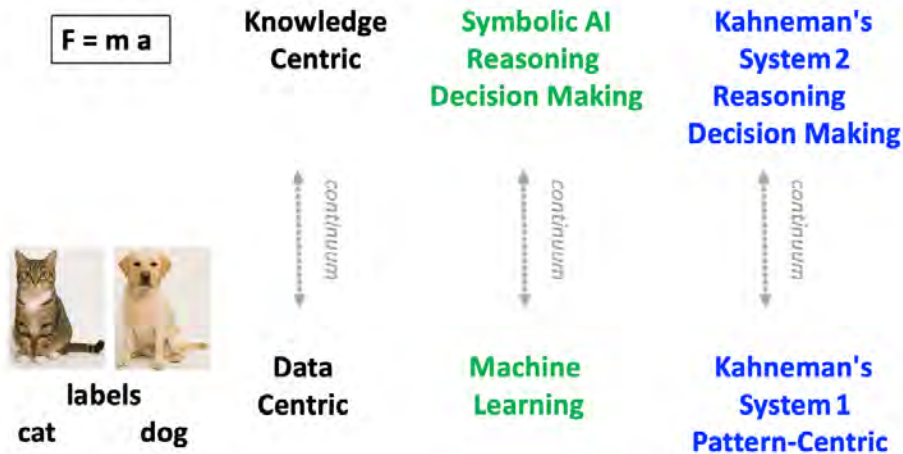
Reasoning is the process of deriving new conclusions from existing knowledge. It makes implicit knowledge explicit and operates on representations of knowledge, such as symbols, graphs, equations, or rules, following principled steps that preserve truth and quantify uncertainty. Reasoning techniques are central to System 2–style cognition, grounded in first principles, rigorous inference, knowledge synthesis, and hypothesis generation.

While data-driven approaches map inputs to outputs through statistical regularities, reasoning manipulates knowledge to derive new facts or hypotheses. Together, they form a continuum: from correlational mapping to explicit knowledge manipulation (Figure 1). At the data-driven end, models uncover statistical regularities and largely interpolate within the training distribution. Moving toward reasoning, outputs gain structure, domain rules are enforced, and causal dependencies become explicit and testable. Scientific and sustainability problems usually sit toward the reasoning side, where scarce data and rich prior knowledge call for hybrids that learn from data while reasoning about prior knowledge.

From early symbolic systems – such as the Logic Theorist for theorem proving; the General Problem Solver I (GPS-I) for understanding the processes required for human problem solving; and DENDRAL, which applied rule-based reasoning to analyze chemical data and infer molecular structures (one of the first AI systems for a scientific domain) – to modern neurosymbolic and probabilistic methods, AI has relied on reasoning and search to solve complex problems and guide discovery.⁸ By integrating data-centric learning with logic, constraints, optimization, and symbolic representations, artificial intelligence can move toward genuine scientific insight. This shift to reasoning requires more explicit representations, greater use of prior knowledge and causal models, deeper planning, and the ability to verify rather than merely score solutions. In practice, one moves along this continuum by making (latent) states interpretable, encoding domain constraints and rules, adding search or external tools, training for consistency, and checking results with separate methods for validation.

In recent years, an emergent line of work has sought to explicitly integrate scientific knowledge, such as physics-based models and governing equations, into

Figure 1
Data Versus Knowledge: Continuums from Correlational Mapping to Explicit Knowledge Manipulation



Source : Illustration provided by the author.

machine-learning frameworks, moving beyond purely data-driven approaches. These efforts encode prior knowledge through modified objectives, architectural constraints, or hybrid combinations of neural and process-based models, reflecting an important shift toward incorporating domain structure into learning systems. At the same time, this integration remains challenging: such models can be difficult to train, do not always fully satisfy underlying scientific laws, and often remain opaque. As a result, their ability to support interpretation and to yield new scientific insight remains, for now, an open challenge, one that may require tighter integration with symbolic and reasoning-based approaches.⁹

Large language models (LLMs), embodied by ChatGPT, Claude, Gemini, and DeepSeek, increasingly pair learning with reasoning and search techniques.¹⁰ Examples include reinforcement learning to incorporate human feedback, as well as chain-of-thought (CoT), tree-of-thoughts (ToT), and graph-of-thoughts (GoT) frameworks for multiple step-by-step inference (a form of “what-if” exploration), to cover large search spaces of possible responses.¹¹ These developments suggest that advanced generative AI and LLMs are moving beyond responses purely based on the learned language modeling toward more deliberate, controlled reasoning and search. A complementary line uses LLMs to evolve algorithms under constraints. The models generate candidate code or proof steps. Verifiers and guardrails

such as tests, invariants, proof checkers, and resource budgets eliminate invalid options. An evaluation loop measures solution rate, runtime, and memory use to identify improvements. Search layers include population-based evolution, best-first or beam search over edits, Monte Carlo tree search or reinforcement learning for sequential choices, and bandit or Bayesian tuning, accepting only population variants that remain correct and faster. This template spans systems for algorithm discovery (AlphaEvolve), formal reasoning (AlphaProof), and the design of logical reasoning engines (SAT solvers).¹²

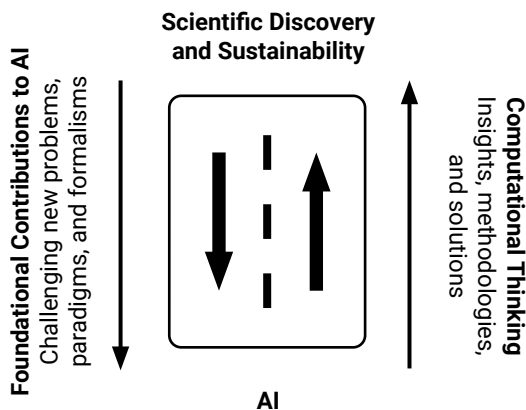
Artificial intelligence for science is a two-way street. This metaphor stems from the core computer science notion of problem reduction, which amplifies the effectiveness of computational methods: Problems that may appear unrelated can often be solved using the same underlying computational approach.¹³ The generality of computational models thus enables the transference of methodologies across domains, underpinning the two-way street shown in Figure 2. By developing general approaches to tackle scientific challenges in sustainability, we not only solve domain-specific problems, but we can also advance artificial intelligence through the creation of general, broadly applicable methodologies. I illustrate the two-way street metaphor with several examples from my lab's work at Cornell involving crystal-structure phase mapping and Deep Reasoning Networks.

AI → Science: Crystal-structure phase mapping is a central challenge in materials science, requiring inference of complex crystal structures from X-ray diffraction (XRD) patterns. A rich body of prior knowledge exists about chemical systems and X-ray diffraction, including thermodynamic principles and Bragg's law, which describes how X-rays are diffracted by atomic planes in a crystal. We developed an approach to automate crystal-structure phase mapping that seamlessly combines learning from diffraction data with knowledge-centric reasoning about known crystal phases and the diffraction process, resulting in a significant improvement in the state-of-the-art of crystalline phase mapping.¹⁴ This work exemplifies how AI techniques can be used to accelerate the scientific discovery process.

Science → AI: Solving crystal-structure phase mapping from X-ray diffraction required embedding scientific laws and constraints within a deep learning framework. This led to our development of the Deep Reasoning Networks (DRNets) framework: an interpretable deep architecture that integrates domain knowledge and enables differentiable reasoning.¹⁵ DRNets combine an encoder that constructs a structured latent space, a reasoning module that enforces domain rules, and a decoder that injects background knowledge. During training, the desired interpretation emerges in the latent space, guided by a combination of unlabeled data with scientific domain knowledge and constraints. The modular design enables the integration of knowledge-centric reasoning with data-centric learning, adaptable to a wide range of scientific and engineering tasks.

Figure 2

The Two-Way Street Between AI and Scientific Discovery



AI enables new approaches for scientific discovery to address challenges in sustainability, and scientific sustainability challenges drive AI advancements. Source: Illustration provided by the author.

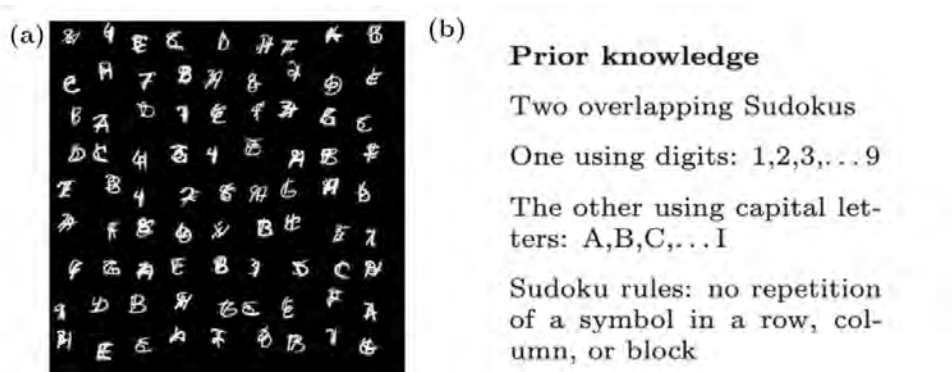
AI → Science (again): The generality of DRNets enables a range of other applications. For instance, the DMVP-DRNets variant (short for Deep Multivariate Probit Deep Reasoning Networks) predicts joint-species distributions, where it is crucial to capture the species' environmental preferences and interactions among and/or between species. The same framework has also been used for mapping the Amazon land cover, multi-object detection, and soil-organic-carbon modeling.¹⁶ Overall, the two-way street metaphor underscores a fundamental aspect of work on AI for scientific discovery, resulting in both novel AI formalisms and accelerated scientific exploration.

Deep Reasoning Networks embed reasoning about prior knowledge directly within deep neural architectures.¹⁷ They seamlessly incorporate knowledge-centric reasoning with data-centric learning through a structured, interpretable latent space that captures domain semantics – enabling more accurate, physics-informed, end-to-end solutions. Originally developed for crystal-structure phase mapping, a core challenge in materials discovery, DRNets extend naturally to other domains. To build intuition about the framework, we begin with a game example (Figure 3) before turning to scientific applications.

Can you make sense of the symbols in Figure 3 (a)? This figure presents a puzzle mixing digits and letters from two overlapping Sudoku games: nine-by-nine grids where each cell must contain a symbol from a given set with no repeats in any row, column, or three-by-three block. Even knowing that the figure represents two

Figure 3

Example of a Demixing Task Based on the Game of Sudoku



(a) shows the mystery pattern of overlapping symbols and (b) is the background knowledge.
Source: Illustration provided by the author.

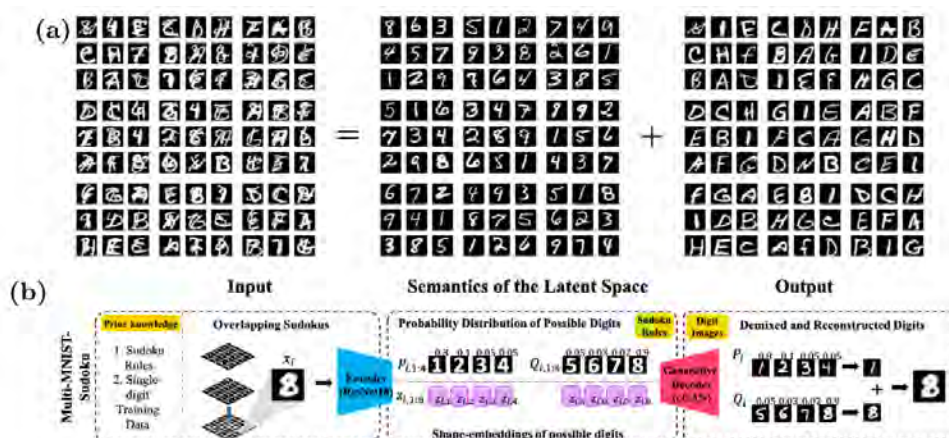
overlapping Sudokus, one using digits (1–9) and the other using letters (A–I), disentangling them is nontrivial. This task, called Multi-MNIST-Sudoku, formalizes demixing overlapping handwritten puzzles.¹⁸ Traditional combinatorial solvers, such as SAT solvers that aim to resolve the Boolean satisfiability problem, excel at pure symbolic tasks but cannot process noisy image data, motivating the development of hybrids that integrate perception and reasoning.¹⁹

DRNets emulate human problem-solving by combining visual evidence with symbolic rules in an end-to-end deep learning framework. An encoder maps each Sudoku cell image to an interpretable latent representation of symbol probabilities and shape features. A conditional generative adversarial network (GAN), informed by single-symbol prototypes, reconstructs digits and letters, while a reasoning module (an LSTM, or long short-term memory network, over the Sudoku constraint graph), coupled with differentiable reasoning constraints, enforces row, column, and block consistency.²⁰ Because Sudoku rules are discrete, DRNets use differentiable, entropy-based continuous relaxations to encode the Sudoku constraints, which are optimized jointly with image reconstruction losses via constraint-aware stochastic gradient descent.

Trained only through self-supervision from relaxed Sudoku rules, without mixed-symbol labels, DRNets outperform supervised baselines, improving both symbol recognition and puzzle completion. Figure 4 (a) shows the demixed pair of Sudokus; Figure 4 (b) gives the DRNets architecture for this task. The knowledge constraints as listed in Figure 3 (b) are enforced through the interpretable latent

Figure 4

Demixed Sudoku Puzzle and the DRNets Framework for the Demixing Task



(a) shows the process of demixing Sudokus and (b) is the DRNets framework and the semantics of the latent space. Source: (a) Illustration provided by the author; and (b) illustration provided by the author using a modified version of Figure 2 (a) in Di Chen, Yiwei Bai, Sebastian Ament, et al., “Automating Crystal-Structure Phase Mapping by Combining Deep Learning with Constraint Reasoning,” *Nature Machine Intelligence* 3 (9) (2021): 812–822, <https://doi.org/10.1038/s42256-021-00384-1>.

space. Interestingly, the correct semantics of the embedded space emerges through self-supervised training. Moreover, the information derived from the image data and from the constraints complement and amplify each other in our integrated framework.

We now turn to scientific applications of the DRNets framework in materials science and sustainability settings. Sustainable materials and processes are crucial to meeting some of the most pressing challenges in energy generation, transportation, and consumption, as well as broader sustainability issues. Long-term progress will rely on breakthrough innovations in materials science, including the development of new materials and processes for more efficient renewable energy systems and for the reduction of carbon dioxide emissions. Here, I highlight our lab’s work in collaboration with materials scientists at Cornell University and at the California Institute of Technology on advancing AI to accelerate materials discovery for renewable energy. Our knowledge-centric AI reasoning approaches are grounded in self-supervision, leveraging the rich scientific knowledge in materials science.

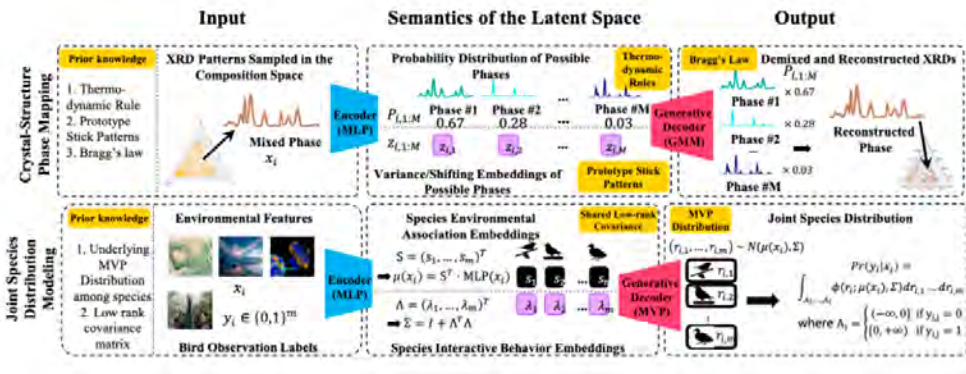
In search of renewable materials for solar fuels, high-throughput experimental settings enable the generation of thousands of materials samples and the analysis of their properties. A major bottleneck in this process, however, is inferring the crystal structure of the samples from high-intensity X-ray diffraction data. Automating crystal-structure phase mapping is therefore essential to speeding up the discovery cycle.

Crystal-structure phase mapping is a demixing task, analogous to Sudoku demixing: Each XRD pattern combines unknown proportions of “pure” phase signals, constrained by the chemistry of the full system. Phase prototypes appear as stick patterns specifying Bragg-peak locations (the positions of strongest X-ray diffraction peaks), analogous to prototypical digits or letters. While simple datasets may be solved by supervised or unsupervised methods, phase mapping is NP-hard, among the most difficult problems to solve computationally, and especially difficult in ternary or higher composition spaces, where alloying causes peak locations to shift and overlap. Addressing this challenge requires reasoning about crystal structures, thermodynamics, and the diffraction process.

Physically and Symmetry Informed DRNets (PSI-DRNets) for crystal-structure phase mapping jointly demix and reconstruct XRD patterns using a graph neural network encoder that produces a structured, interpretable latent space, and a generative decoder that integrates Bragg’s law – linking X-ray wavelength, diffraction angle, and crystal plane spacing – with a Gaussian mixture model (GMM) to represent peak shapes.²¹ The decoder incorporates prior knowledge from previously characterized XRD prototype patterns, which are ideal, noise-free profiles from resources such as the Inorganic Crystal Structure Database, used to guide reconstruction.²² The interpretable latent space encodes phase probabilities and lattice parameters, while thermodynamic and symmetry constraints are enforced via differentiable entropy-based functions with a batched sampling strategy to handle combinatorial complexity across multiple samples (see the top panel of Figure 5 that shows a simplified version of the latent space). A hybrid loss function balances reconstruction fidelity and constraint satisfaction, allowing us to solve phase-mapping instances that had resisted both traditional algorithms and expert manual analysis.

The DRNets framework outperforms state-of-the-art methods and, more important, succeeds where human experts could not: for example, solving the previously unsolved Bi – Cu – V oxide system, a breakthrough that enabled the discovery of solar fuels materials by revealing a three-phase mixture – alloy variants of BiVO_4 , Cu_2BiVO_6 , and $\text{Cu}_3(\text{VO}_4)_2$ – that outperformed monoclinic BiVO_4 in photoactivity, overturning the assumption that phase mixtures inherently reduce performance.²³ These results illustrate how combining deep learning with first-principles, knowledge-centric reasoning can automate data interpretation and significantly accelerate scientific discovery.

Figure 5
Semantics of the Interpretable Latent Space of the DRNets Framework for Crystal-Structure Phase Mapping and Joint Species–Distribution Modeling



(Top) In crystal-structure phase mapping, DRNets encode the input XRD patterns x_i into $P_{i,1:M}$ and $z_{i,1:M}$, which denote the probability distribution and the variance/shifting embeddings of M possible phases. This is a simplified version that does not depict the crystal-structure lattice parameters, also captured by the latent space. The generative decoder (GMM) uses $z_{i,1:M}$ to generate the decomposed phases and reconstruct the original XRD using the phase probability distribution $P_{i,1:M}$. (Bottom) In joint species–distribution modeling, DRNets extract the high-level environmental features $MLP(x_i)$ from the raw input x_i and encode the species environmental association embeddings s_j and the interactive behavior embeddings λ_j for each species j . The generative decoder (MVP) uses s_j , $MLP(x_i)$, and λ_j to compute the corresponding $\mu(x_i)$ and Σ for the latent multivariate Gaussian random variables $r_{i,j}$ and maximizes the likelihood of the species observation y_i given x_i under the multivariate probit distribution. Source: Illustration provided by the author. The top panel is a modified version of Figure 2 (b) in Di Chen, Yiwei Bai, Sebastian Ament, et al., “Automating Crystal-Structure Phase Mapping by Combining Deep Learning with Constraint Reasoning,” *Nature Machine Intelligence* 3 (9) (2021): 812–822, <https://doi.org/10.1038/s42256-021-00384-1>.

Biodiversity represents the variety of life on Earth and is essential to human well-being, supporting healthy environments and providing natural resources such as food, water, and medicine. It also offers important economic, cultural, and aesthetic benefits. High biodiversity strengthens ecosystem resilience, helping systems recover from disturbances such as climate change and other extreme events.

Understanding how species are distributed across space and time is critical for guiding effective conservation efforts. Inspired by joint species–distribution models and powered by over nine million records of eBird, an online database and citizen-science project for bird observations run by the Cornell Laboratory of Ornithology, we applied the DRNets framework to model year-round spatiotemporal distributions, species–environment associations, and interspecies interactions for

five hundred North American bird species, at a scale surpassing prior joint species–distribution models.²⁴ Scientific ecological knowledge enhances interpretability and improves joint-species prediction by constraining models to realistic regimes not fully captured by labeled data.

The bottom panel of Figure 5 depicts the DMVP-DRNets architecture: an end-to-end model with a semantically structured latent space that imposes an ecology-informed low-rank constraint on the shared covariance and that uses efficient sampling to estimate species interactions explicitly, paired with a multivariate probit (MVP) decoder based on a multivariate Gaussian model.²⁵ Unlike many deep models (such as variational autoencoders, or VAEs) that assume diagonal covariance and thus independence, an ecologically unrealistic simplification, DMVP-DRNets capture cross-species dependence directly.²⁶

The DMVP-DRNets framework reveals both bird habitat associations and species co-occurrence patterns, producing fine-grained richness maps for around five hundred species, the entire North American avifauna, critical for conservation planning and management. By encoding prior ecological knowledge and modeling interactions explicitly, it improves interpretability and robustness, extending for the first time joint-species modeling to a continental scale.

As another example, the Scientifically-Interpretable Reasoning Network (SciReN) architecture, which models the soil-organic-carbon cycle, is an end-to-end framework with a fully interpretable encoder mapping inputs to an interpretable latent space of over twenty biogeochemical parameters and driving a differentiable decoder based on the Community Land Model version 5 (CLM5), a process-based climate model.²⁷ The framework’s predictive power matches or exceeds that of prior approaches and, crucially, enables interpretation of the underlying biogeochemical parameters.

In summary, the DRNets framework seamlessly integrates reasoning and deep learning: An encoder builds an interpretable structured latent space for integrating prior knowledge, a reasoning module enforces domain rules, and a decoder injects background insights. This architecture (of encoder + interpretable latent space + reasoning module + decoder) combines knowledge-centric reasoning with data-centric learning: it leverages from a true integration of data-driven and knowledge-centric approaches, and its modular design makes it broadly applicable across domains.

Multi-agent knowledge-centric frameworks will play a central role in AI for scientific discovery, as highly specialized, (semi-)autonomous agents learn, reason, and collaborate while interacting with digital and physical environments.²⁸ Our work on SARA, the Scientific Autonomous Reasoning Agent, embodies this vision for materials discovery: a robot scientist that operationalizes the scientific method through hypothesis formulation, experimental design,

planning, execution, interpretation, and knowledge generation.²⁹ SARA already conducts closed-loop experiments using active learning to control laser-spike-annealing synthesis of metastable oxide materials, collect optical and X-ray diffraction data, and map the underlying crystal structures. In SARA's ecosystem, different agents apply various elements of the scientific method and coordinate so that work flows seamlessly from data and knowledge to insight, from conjecture to experiment. The PSI-DRNets component for crystal-structure phase mapping, described earlier, is one of SARA's agents. Work with A-Lab, another autonomous agent, has shown that integrating literature-trained planning, active learning, robotics, and rapid characterization can automate key steps in inorganic materials discovery.³⁰ Complementing this automation, recent work on scientific reasoning for chemistry suggests that domain-tuned models can articulate step-by-step chemical reasoning and produce structured molecular representations, highlighting the role of explicit reasoning in autonomous discovery.³¹

To make knowledge-centric SciAIs effective, they must communicate using clear and concise representations that both human scientists and other AI agents can interpret and exploit. Learning should proceed not only from data but also from prior scientific knowledge, including physical laws and domain constraints that act as training signals in data-limited regimes. Interpretable latent spaces should align with scientific semantics, allowing SciAIs to incorporate and effectively exploit prior knowledge. Reasoning can leverage problem structure. For example, agents can identify symmetries and invariants, surface the constraints that matter, and recognize when a task can be translated, reduced, or reformulated as one that another agent already knows how to solve efficiently. In this way, the art of modeling and problem reduction becomes a general capability shared across agents and domains.

As autonomy increases, these SciAIs must couple inference with action in self-driving laboratories. A self-driving lab links planning, measurement, and control so that it not only interprets results but also decides what to do next. It selects the next sample, tunes process parameters, schedules measurements, and updates hypotheses as data arrive, enabling disciplined, iterative cycles that converge on reliable knowledge. Decisions are guided by active learning based on the value of information, by safety and feasibility constraints, and by an understanding of how errors propagate through models.

Scientific progress is rarely defined by a single objective; SciAIs and self-driving labs must reason about trade-offs. In particular, AI systems should reveal the Pareto frontier among multiple objectives such as performance, robustness, interpretability, cost, and sustainability criteria; help scientists understand the shape of that frontier; and quantify the consequences of moving along it.³² Trade-off analysis makes explicit the values behind each choice. For example, in a sustainability setting concerning strategic hydropower planning, my lab, in collaboration with a highly international and interdisciplinary group of scientists, demonstrated how Pareto

multi-objective optimization can balance energy generation with the reduction of adverse impacts on people and nature.³³

Specialized SciAIs coordinate through shared abstractions and verified messages. Deeply specialized agents can be more effective since their knowledge and reasoning can be precisely tuned to the particular structure of a specific domain. One agent finds the right problem formulation, for example, a materials question, another finds the best suitable algorithm to solve it under constraints, and a third maps the solution back into physical parameters and predicted outcomes. The resulting chain can be checked and validated end-to-end, with provenance and assumptions carried forward rather than hidden – enabling transparency and reproducibility, preconditions for scientific autonomy.

Science advances by leveraging accumulated knowledge, as beautifully captured by Isaac Newton, who, in 1675, wrote in a letter to Robert Hooke, “If I have seen further, it is by standing on the shoulders of giants.”³⁴ Knowledge-centric AI and reasoning from first principles offer significant new opportunities for accelerating scientific exploration and discovery. AI for science dates to the early days of the 1960s, with DENDRAL, one of the first expert systems, developed at Stanford University under the leadership of Edward Feigenbaum, with a team that included Nobel Prize-winning molecular biologist Joshua Lederberg and pharmaceutical chemist Carl Djerassi. DENDRAL used rule-based reasoning to analyze chemical data and infer molecular structures. Over the past decade, AI has ridden a data-centric wave powered by deep learning and large language models. Here, I argue for injecting knowledge-centric reasoning into AI systems and that genuine discovery requires first-principles reasoning over background and scientific knowledge, tightly integrated with data-centric learning. Our work supports this thesis through a focus on computational sustainability: a two-way street in which AI methods address environmental and societal challenges, and those challenges, in turn, spur new methodologies that advance AI. By tackling sustainability-driven scientific tasks, including pattern demixing for phase mapping in materials science, bioacoustic call identification, joint species–distribution modeling, materials property prediction, land-cover mapping, object detection, and soil-organic-carbon modeling, we have developed general methodologies that transfer across domains.

Looking ahead, scientific progress is poised to be shaped by communities of knowledge-centric scientific AI agents that learn, reason, and collaborate across domains, close the loop between inference and action in (semi-)autonomous laboratories, and navigate trade-offs transparently. In this vision, artificial intelligence becomes a valuable scientific collaborator.

AUTHOR'S NOTE

This work was partially supported by an AI2050 Senior Fellowship from Schmidt Sciences, the National Science Foundation, the National Institute of Food and Agriculture, and the Air Force Office of Scientific Research.

ABOUT THE AUTHOR

Carla P. Gomes is the Ronald C. and Antonia V. Nielsen Professor of Computing and Information Science at Cornell University, Director of the Institute for Computational Sustainability at Cornell, and Codirector of the Cornell University AI for Science Institute. She is a founder of the field of computational sustainability and the author or coauthor of over two hundred publications. She has published in such journals as *Nature*, *Science*, and *Advances in Neural Information Processing Systems*.

ENDNOTES

- ¹ David Silver, Aja Huang, Chris J. Maddison, et al., “Mastering the Game of Go with Deep Neural Networks and Tree Search,” *Nature* 529 (7587) (2016): 484–489, <https://doi.org/10.1038/nature16961>; Josh Abramson, Jonas Adler, Jack Dunger, et al., “Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3,” *Nature* 630 (8016) (2024): 493–500, <https://doi.org/10.1038/s41586-024-07487-w>; OpenAI, Josh Achiam, Steven Adler, et al., “GPT-4 Technical Report,” arXiv (2023), <https://doi.org/10.48550/arXiv.2303.08774>; and Jared Willard, Xiaowei Jia, Shaoming Xu, et al., “Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems,” *ACM Computing Surveys* 55 (4) (2022): 1–37.
- ² Ben Goertzel, “Artificial General Intelligence: Concept, State of the Art, and Future Prospects,” *Journal of Artificial General Intelligence* 5 (1) (2014): 1–48, <https://doi.org/10.2478/jagi-2014-0001>.
- ³ United Nations General Assembly, *Transforming Our World: The 2030 Agenda for Sustainable Development*, A/RES/70/1 (September 25, 2015), <https://docs.un.org/en/A/RES/70/1>.
- ⁴ Carla Gomes, Thomas Dietterich, Christopher Barrett, et al., “Computational Sustainability: Computing for a Better World and a Sustainable Future,” *Communications of the ACM* 62 (9) (2019): 56–65, <https://doi.org/10.1145/3339399>; and Carla P. Gomes, “Computational Sustainability: Computational Methods for a Sustainable Environment, Economy, and Society,” *The Bridge* 39 (4) (2009), <https://www.nae.edu/Publications/Bridge/17281/17286.aspx>.
- ⁵ Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, “Deep Learning,” *Nature* 521 (7553) (2015): 436–444, <https://doi.org/10.1038/nature14539>; and Ashish Vaswani, Noam Shazeer, Niki Parmar, et al., “Attention Is All You Need,” in *Advances in Neural Information Processing Systems* 30 (NIPS 2017), ed. Isabelle Guyon, Ulrike Von Luxburg, Samy Bengio, et al. (Neural Information Processing Systems Foundation, 2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
- ⁶ Daniel Kahneman, *Thinking, Fast and Slow* (Farrar, Straus and Giroux, 2011), 499.

- ⁷ Hanchen Wang, Tianfan Fu, Yuanqi Du, et al., “Scientific Discovery in the Age of Artificial Intelligence,” *Nature* 620 (7972) (2023): 47–60, <https://doi.org/10.1038/s41586-023-06221-2>.
- ⁸ Allen Newell and Herbert Simon, “The Logic Theory Machine—A Complex Information Processing System,” *IRE Transactions on Information Theory* 2 (3) (1956): 61–79, <https://doi.org/10.1109/TIT.1956.1056797>; Allen Newell, J. C. Shaw, and Herbert A. Simon, “Report on a General Problem-Solving Program,” in *Information Processing: Proceedings of the International Conference on Information Processing* (UNESCO, 1960), 256–264, http://bitsavers.informatik.uni-stuttgart.de/pdf/rand/ipl/P-1584_Report_On_A_General_Problem-Solving_Program_Feb59.pdf; Robert K. Lindsay, Bruce G. Buchanan, Edward A. Feigenbaum, and Joshua Lederberg, “DENDRAL: A Case Study of the First Expert System for Scientific Hypothesis Formation,” *Artificial Intelligence* 61 (2) (1993): 209–261, [https://doi.org/10.1016/0004-3702\(93\)90068-M](https://doi.org/10.1016/0004-3702(93)90068-M); and Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. (Pearson, 2021), 1115.
- ⁹ Willard, Jia, Xu, et al., “Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems”; George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, et al., “Physics-Informed Machine Learning,” *Nature Reviews Physics* 3 (6) (2021): 422–440; Maziar Raissi, Paris Perdikaris, and George Em Karniadakis, “Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations,” *Journal of Computational Physics* 378 (2019): 686–707; and Aditi S. Krishnapriyan, Amir Gholami, Shandian Zhe, et al., “Characterizing Possible Failure Modes in Physics-Informed Neural Networks,” *Advances in Neural Information Processing Systems* 34 (*NeurIPS* 2021), ed. Marc’Aurelio Ranzato, Alina Beygelzimer, Yann Dauphin, et al. (Curran Associates, Inc., 2021), 26548–26560.
- ¹⁰ OpenAI, Achiam, Adler, et al., “GPT-4 Technical Report”; Anthropic, “The Claude 3 Model Family: Opus, Sonnet, Haiku,” https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c3b80b7bbc618857627/Model_Card_Claude_3.pdf (accessed March 9, 2026); Google Gemini Team, “Gemini: A Family of Highly Capable Multimodal Models,” arXiv (2023), <https://doi.org/10.48550/arXiv.2312.11805>; and Daya Guo, Dejian Yang, Haowei Zhang, et al., “DeepSeek-R1 Incentivizes Reasoning in LLMs Through Reinforcement Learning,” *Nature* 645 (8081) (2025): 633–638, <https://doi.org/10.1038/s41586-025-09422-z>.
- ¹¹ Richard S. Sutton and Andrew G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. (MIT Press, 2018), 526; Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” in *Advances in Neural Information Processing Systems* 35 (*NeurIPS* 2022), ed. Sanmi Koyejo, Shakir Mohamed, Alekh Agarwal, et al. (Curran Associates, Inc., 2022), 24824–24837, <https://dl.acm.org/doi/10.5555/3600270.3602070>; Shunyu Yao, Dian Yu, Jeffrey Zhao, et al., “Tree of Thoughts: Deliberate Problem Solving with Large Language Models,” in *Advances in Neural Information Processing Systems* 36 (*NeurIPS* 2023), ed. Alice Oh, Tristan Naumann, Amir Globerson, et al. (Curran Associates, Inc., 2023), 11809–11822, <https://dl.acm.org/doi/abs/10.5555/3666122.3666639>; and Maciej Besta, Nils Blach, Ales Kubicek, et al., “Graph of Thoughts: Solving Elaborate Problems with Large Language Models,” in *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, ed. Michael Wooldridge, Jennifer Dy, and Sriraam Natarajan (AAAI Press, 2024), 17682–17690, <https://doi.org/10.1609/aaai.v38i16.29720>.
- ¹² Alexander Novikov, Ngán Vŭ, Marvin Eisenberger, et al. “AlphaEvolve: A Coding Agent for Scientific and Algorithmic Discovery,” arXiv (2025), <https://doi.org/10.48550/arXiv.2506.13131>; Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, et al., “Mathematical Discoveries from Program Search with Large Language

- Models,” *Nature* 625 (7995) (2024): 468–475, <https://doi.org/10.1038/s41586-023-06924-6>; and Cunxi Yu, Rongjian Liang, Chia-Tung Ho, and Haoxing Ren, “Autonomous Code Evolution Meets NP-Completeness,” arXiv (2025), <https://doi.org/10.48550/arXiv.2509.07367>.
- ¹³ Richard M. Karp, “Reducibility Among Combinatorial Problems,” in *50 Years of Integer Programming, 1958–2008: From the Early Years to the State-of-the-Art*, 1st edition, ed. Michael Jünger, Thomas M. Liebling, Denis Naddef, et al. (Springer Berlin Heidelberg, 2010), 219–241.
- ¹⁴ Di Chen, Yiwei Bai, Sebastian Ament, et al., “Automating Crystal-Structure Phase Mapping by Combining Deep Learning with Constraint Reasoning,” *Nature Machine Intelligence* 3 (9) (2021): 812–822, <https://doi.org/10.1038/s42256-021-00384-1>.
- ¹⁵ Di Chen, Yiwei Bai, Wenting Zhao, et al., “Deep Reasoning Networks for Unsupervised Pattern Demixing with Constraint Reasoning,” in *ICML’20: Proceedings of the 37th International Conference on Machine Learning*, ed. Hal Daumé and Aarti Singh (JMLR, 2020), 1500–1509.
- ¹⁶ Courtney L. Davis, Yiwei Bai, Di Chen, et al., “Deep Learning with Citizen Science Data Enables Estimation of Species Diversity and Composition at Continental Extents,” *Ecology* 104 (12) (2023), <https://doi.org/10.1002/ecy.4175>; Chen, Bai, Zhao, et al., “Deep Reasoning Networks for Unsupervised Pattern Demixing with Constraint Reasoning”; and Joshua Fan, Haodi Xu, Feng Tao, et al., “Scientifically-Interpretable Reasoning Network (SciReN): Discovering Hidden Relationships in the Carbon Cycle and Beyond,” arXiv (2025), <https://doi.org/10.48550/arXiv.2506.14054>.
- ¹⁷ Chen, Bai, Ament, et al., “Automating Crystal-Structure Phase Mapping by Combining Deep Learning with Constraint Reasoning.”
- ¹⁸ Chen, Bai, Zhao, et al., “Deep Reasoning Networks for Unsupervised Pattern Demixing with Constraint Reasoning.”
- ¹⁹ Vijay Ganesh and Moshe Y. Vardi, “On the Unreasonable Effectiveness of SAT Solvers,” in *Beyond the Worst-Case Analysis of Algorithms*, ed. Tim Roughgarden (Cambridge University Press, 2020).
- ²⁰ Mehdi Mirza and Simon Osindero, “Conditional Generative Adversarial Nets,” arXiv (2014), <https://doi.org/10.48550/arXiv.1411.1784>; and Sepp Hochreiter and Jürgen Schmidhuber, “Long Short-Term Memory,” *Neural Computation* 9 (8) (1997): 1735–1780, <https://doi.org/10.1162/neco.1997.9.8.1735>.
- ²¹ Yimeng Min, Ming-Chiang Chang, Shufeng Kong, et al., “Physically Informed Graph-Based Deep Reasoning Net for Efficient Combinatorial Phase Mapping,” in *2023 International Conference on Machine Learning and Applications (ICMLA)* (Curran Associates, Inc., 2023), 392–399, <https://doi.org/10.1109/ICMLA58977.2023.00061>.
- ²² FIZ Karlsruhe–Leibniz Institute for Information Infrastructure, “About ICSD,” <https://icsd.products.fiz-karlsruhe.de/en/about/about-icsd> (accessed March 11, 2026).
- ²³ Chen, Bai, Ament, et al., “Automating Crystal-Structure Phase Mapping by Combining Deep Learning with Constraint Reasoning.”
- ²⁴ Di Chen, Yexiang Xue, Daniel Fink, et al., “Deep Multi-Species Embedding,” in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, ed. Carles Sierra (International Joint Conferences on Artificial Intelligence, 2017), 3639–3646, <https://doi.org/10.24963/ijcai.2017/509>.
- ²⁵ Davis, Bai, Chen, et al., “Deep Learning with Citizen Science Data Enables Estimation of Species Diversity and Composition at Continental Extents.”

- ²⁶ Diederik P. Kingma and Max Welling, “Auto-Encoding Variational Bayes,” arXiv (2013), <https://doi.org/10.48550/arXiv.1312.6114>.
- ²⁷ Fan, Xu, Tao, et al., “Scientifically-Interpretable Reasoning Network (SciREN)” ; and David M. Lawrence, Rosie A. Fisher, Charles D. Koven, et al., “The Community Land Model Version 5: Description of New Features, Benchmarking, and Impact of Forcing Uncertainty,” *Journal of Advances in Modeling Earth Systems* 11 (12) (2019): 4245–4287, <https://doi.org/10.1029/2018MS001583>.
- ²⁸ Hongliang Xin, John R. Kitchin, and Heather J. Kulik, “Towards Agentic Science for Advancing Scientific Discovery,” *Nature Machine Intelligence* 7 (9) (2025): 1373–1375, <https://doi.org/10.1038/s42256-025-01110-x>.
- ²⁹ Sebastian Ament, Maximilian Amsler, Duncan R. Sutherland, et al., “Autonomous Materials Synthesis via Hierarchical Active Learning of Nonequilibrium Phase Diagrams,” *Science Advances* 7 (51) (2021), <https://doi.org/10.1126/sciadv.abg4930>.
- ³⁰ Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei, et al., “An Autonomous Laboratory for the Accelerated Synthesis of Inorganic Materials,” *Nature* 624 (7990) (2023): 86–91, <https://doi.org/10.1038/s41586-023-06734-w>.
- ³¹ Siddharth M. Narayanan, James D. Braza, Ryan-Rhys Griffiths, et al., “Training a Scientific Reasoning Model for Chemistry,” arXiv (2025), <https://doi.org/10.48550/arXiv.2506.17238>.
- ³² Arun Mannodi-Kanakkithodi, Ghanshyam Pilania, Rampi Ramprasad, et al., “Multi-Objective Optimization Techniques to Design the Pareto Front of Organic Dielectric Polymers,” *Computational Materials Science* 125 (2016): 92–99, <https://doi.org/10.1016/j.commatsci.2016.08.018>.
- ³³ Alexander S. Flecker, Qinru Shi, Rafael M. Almeida, et al., “Reducing Adverse Impacts of Amazon Hydropower Expansion,” *Science* 375 (6582) (2022): 753–760, <https://doi.org/10.1126/science.abj4017>.
- ³⁴ The Historical Society of Pennsylvania, “Isaac Newton Letter to Robert Hooke, 1675,” <https://digitallibrary.hsp.org/Detail/objects/9792#> (accessed March 6, 2026).

Building an AI Polymath

Shirley Ho

Artificial intelligence has made remarkable strides in natural language processing and image recognition, yet its impact on the natural sciences is fragmented. While specialized models like AlphaFold have revolutionized biology, the scientific enterprise remains siloed, with most foundational models narrowly tailored to specific domains or modalities. In this essay, I advocate for a new class of scientific AI: the polymathic foundation model. Inspired by the intellectual versatility of human polymaths, such a model would integrate diverse data types and disciplinary knowledge across the natural sciences. I argue that building such a model is not only technically feasible but epistemologically necessary. I draw on lessons from existing interdisciplinary successes and outline key challenges: scientific dataset curation, multimodal and multitask learning, verifiable knowledge exchange, and interpretability. I close the essay with a cautiously optimistic roadmap for how such models could transform scientific discovery in the next decade.

Over the past decade, artificial intelligence has redefined the boundaries of what machines can do. Large language models such as GPT-4 and multimodal systems like Gemini or Claude can now perform tasks that once seemed reserved for human intelligence. In narrow scientific domains, models like AlphaFold have cracked decades-old biological puzzles.¹ In astrophysics, AI is accelerating and improving on simulations of the entire universe; scientists investigating the fundamental parameters of the universe are using AI to analyze ever more complex datasets; and researchers are employing AI to find extremely rare cosmic objects.² Beyond the more traditional approaches of AI, astrophysicists and machine learners have built foundation models of astronomy including more than two hundred million celestial objects, allowing astronomers to super-resolve low-resolution images, spectra, or time-series data. The models have learned the underlying properties of cosmic objects, drawing from thirty-nine different ways of looking at them.³ Particle physicists are using AI to make diagnostics of beam trajectories.⁴ In chemistry, labs are relying on AI to find better molecules with particular properties.⁵ However, the promise of a truly general scientific model – one capable of learning, reasoning, and transferring knowledge across the natural sciences – remains unfulfilled. Assisted by AI, we are officially in the age of *fragmented genius*.

This absence reflects a deeper problem: Modern science, despite its empirical rigor, is increasingly fragmented. Each subfield operates within its own conceptual and methodological silo. But nature does not obey disciplinary boundaries. A human, a solar flare, or our evolving planet Earth are not merely phenomena of medicine, heliophysics, or planetary physics alone. They are dynamic, multiscale systems that require integrated, cross-domain understanding. What we lack is not computing power or data, but a new kind of epistemic infrastructure: a generalist AI model that can traverse the scientific landscape with polymathic agility. This essay presents a vision for such a model and presents the methodological, infrastructural, and philosophical considerations necessary for its realization.

The analogy with human polymaths is instructive. Thinkers like Leonardo da Vinci, Issac Newton, and Benjamin Franklin did not operate within strict disciplinary lines. Their insights emerged from recognizing deep patterns across domains. A polymathic model should similarly be able to both leverage data from plasma physics on the ground to improve understanding of cosmic dynamics and draw on fluid dynamics to reinterpret climate simulations. However, building a polymathic model will require ingesting scientific data that go far beyond what typical large language models (LLMs) can handle. Most current foundation models in AI are trained on massive amounts of unstructured text, images, and videos, allowing them to perform a wide range of multimodal tasks. However, to make real discoveries, we need to go beyond the scientific literature and code repositories. We need structured and unstructured scientific data extracted from telescopes, temperature sensors, medical imagers, pressure sensors, and more. These scientific data usually go beyond text or RGB images and videos. They include time series, simulation outputs, graphs, images, and structured metadata with extreme ranges. A polymathic model should be able to process this vast range of data modalities while preserving their complexities and specificities. A polymathic model also requires scientific reasoning, which demands more than just prediction. It requires interpretability, causality, and the ability to abstract across disciplines and scales. A model that predicts solar activity must also provide uncertainties around its predictions; it must explain why certain features matter and how they relate to physical theories. Our physical theory of the sun should also explain activity of faraway stars, or the interactions of atoms in a fusion reactor. A generalist model must therefore do more than interpolate patterns; it must *synthesize understanding*.

Current artificial intelligence systems pursued in frontier labs, despite remarkable capabilities, fundamentally lack a grounded understanding of the physical world. Large language and image models excel at manipulating text and image patterns within scientific literature but cannot reason from first principles about physical, chemical, or biological phenomena or observations. Meanwhile, special-

ized scientific models such as AlphaFold, ESM (Evolutionary Scale Modeling), and ChemBERTa possess deep domain knowledge but are usually isolated black boxes. The central challenge we face in our generation is how to create artificial general intelligence (AGI) systems that integrate real scientific reasoning with comprehensive scientific understanding across all domains.

But before we go too far, I should clarify what I mean by AGI in this context. For this essay, I define artificial general intelligence as the intelligence level I would expect from a talented undergraduate, who is able to do many things well but not perfectly and can learn new things quickly.

This represents the missing piece in AGI development: systems that understand not just human language but the fundamental principles governing reality itself. For a lot of us, the goal of creating superhuman artificial intelligence is to create scientific and technological breakthroughs at a rate that is hundreds or millions of times faster than current-day scientific and technological workflows. Current AI models remain fundamentally black boxes. Unlike general language models, where interpretability focuses on linguistic reasoning, scientific models encode complex domain-specific causal relationships, physical laws, and multiscale interactions that are essential for AGI systems to understand and potentially interact with reality.

Scientific inquiry frequently examines identical phenomena through disparate disciplinary lenses: a planetary system, for instance, may be scrutinized via orbital mechanics, atmospheric chemistry, or radiative dynamics. While each vantage point is rigorous, no isolated perspective encapsulates the entirety of the system. In clinical medicine, human health is similarly multilayered, parsed into genetic architecture, proteomic profiles, cellular environments, organ histology, symptomatic manifestations, diagnostic imaging, and longitudinal clinical narratives. Physics formalizes this via effective theories: by abstracting to a specific scale, one can derive a consistent formulation by which the system adheres to emergent laws governing that or higher resolutions.

A polymathic AI model must consequently synthesize and reason across these multiscale, heterogeneous data streams. This necessitates the engineering of shared latent representations capable of encoding both localized structural motifs – such as the stochastic evolution of a fluid vortex – and overarching global constraints, including symmetry groups and fundamental conservation laws. Metadata serve as the essential connective tissue in this architecture. Scientific datasets are permeated with vital contextual parameters – instrumentation specifications, calibration protocols, and foundational physical assumptions – that must be explicitly preserved. One robust strategy involves integrating metadata tokens directly into the model's input stream, empowering the architecture to autonomously learn pre-processing heuristics.

The realization of a polymathic artificial intelligence is not merely a matter of scaling existing architectures but of resolving fundamental tensions between how machines learn and how science is structured. If we are to move beyond the current “standard model” of narrow scientific AI, we must cross several primary conceptual and technical hurdles.

Current scientific data exist in a state of digital feudalism. Unlike the natural language corpora that fueled the large language model revolution, scientific data are characterized by idiosyncratic formats and a lack of unified metadata. To bridge these gaps, we require a Science Data Bank: a curated, cross-disciplinary infrastructure that treats data not as a static resource but as a dynamic graph of physical relationships that can be easily accessed by a wide range of communities.

The rapid maturation of natural language models was catalyzed by meticulously curated corpora such as the Pile, the 886-gigabyte text file used to train many LLMs, and rigorous benchmarking frameworks like GLUE.⁶ But scientific AI currently lacks a comparable foundational infrastructure. Scientific data remain largely fragmented, typically necessitating profound domain expertise for meaningful preprocessing, validation, and interpretive analysis.

A primary obstacle to replicating the linguistic revolution within the sciences by engineering a generalist scientific model is the strategic curation of a massive, multimodal dataset paired with standardized evaluation metrics spanning diverse disciplines. This endeavor presents both a formidable technical challenge and a significant opportunity to investigate knowledge transfer across disparate disciplinary boundaries.

Historically, researchers across various scientific fields have concentrated on isolated seed domains; noteworthy large-scale curation efforts are already underway in astronomy, biology, fluid dynamics, and automotive engineering.⁷ As the “AI for science” ecosystem evolves, we must establish scalable methodologies for dataset expansion. To determine which datasets to prioritize, I propose an expansion strategy governed by two central principles: the anticipated scientific and societal impact of the underlying problem and the prioritization of domains that offer the highest probability of positive knowledge transfer to a few chosen core seed problems. But what does a comprehensive plan for data curation and domain expansion look like? And how can we simultaneously quantify and gain deeper insight into interdisciplinary knowledge transfer?

I advocate for the development of a centralized, expandable infrastructure – the Science Data Bank – designed to house high-fidelity, multimodal scientific datasets. Domains such as fluid dynamics and astrophysics provide an ideal initial substrate; these fields are not only inherently impactful but also share fundamental structural symmetries, such as conservation laws, that facilitate cross-domain generalization. Constructing this Science Data Bank necessitates a sophisticated hybrid strategy: leveraging expert-driven curation for foundational phases fol-

lowed by AI-assisted autonomous discovery. This data infrastructure is as vital to polymathic intelligence as the architecture of the model itself. Development will be divided into three phases:

1. *Small-scale expert-driven data curation and benchmark creation.* As previously examined, several specialized research cohorts have mobilized extensive networks of domain experts to architect high-fidelity initial datasets and corresponding benchmarks. While currently confined to a limited number of disciplines, these repositories already demonstrate the requisite scale and structural diversity to underpin significant foundation models while simultaneously providing high-utility resources for the broader scientific community.
2. *Mid-scale collaboration-driven data curation.* The subsequent phase leverages the collective intelligence of the broader scientific ecosystem through a collaborative framework. In this model, select teams spearhead community-wide initiatives to aggregate and extend existing datasets into novel domains. This expansion can be facilitated through diverse institutional mechanisms, including specialized incubator programs and workshop-driven engagements. Such forums will bridge disparate fields, bringing together experts from adjacent disciplines to identify critical scientific challenges that remain intractable using conventional methodologies. Open calls for proposals can also incentivize external teams to contribute new datasets and benchmarks, an approach increasingly championed by federal agencies like the Department of Energy and National Science Foundation, as evidenced by their recent strategic mandates.⁸
3. *Large-scale automated data curation.* Given the rapid acceleration of large language models and vision-language models, I advocate for a progressive transition toward algorithmic automation to orchestrate dataset discovery and training prioritization. This automated infrastructure is vital for scaling these efforts across the full spectrum of scientific inquiry. The primary methodology involves performing a large-scale semantic synthesis of the existing scientific literature to map interconnected conceptual nodes across disparate topics. By consolidating these data into a comprehensive knowledge graph, we can visualize the current landscape of scientific foundation models, identifying both established territories and critical coverage gaps. This graph-theoretic approach enables the scalable identification of fields linked by shared vocabularies, overlapping data modalities, or analogous analytical requirements. Such structural commonalities serve as potent predictors of knowledge transfer, which can be further refined through associative metrics such as dataset co-occurrence and cross-citation patterns. Complementarily, we can employ active learning strategies to predict mathematically which

novel datasets will yield the highest marginal improvement in model performance. By training models on specific seed domains and defining a robust utility function, we can quantitatively assess the transformative potential of acquiring proposed data streams.

A persistent critique of deep learning or artificial intelligence is its tendency to prioritize statistical correlation over physical causation. For a model to be truly polymathic, one would imagine that its architecture must natively respect the invariants of the natural world: symmetry, conservation, and scale. However, recent developments in large-scale model training suggest that *given enough data*, the fundamental laws of nature can be learned directly from the data, removing the necessity of embedding these symmetries and conservation laws directly within the neural networks themselves. The balance between data quantity and the inductive biases we provide to the AI system *a priori* is one that may surprise us all. I would also suggest explorations toward architectures that can draw inferences through the use of reason across different types of data in different domains. A true AGI should be able to navigate the vast gulf between the quantum behavior of molecules and the macroscopic dynamics of fluid flow with ease.

Cultivating a granular understanding of potential knowledge transfer across heterogeneous datasets and domains is a cornerstone of our expansion strategy. This is a critical requirement for engineering robust, data-driven scientific models. By addressing this problem, we can move beyond conceptual inquiries regarding multidisciplinary to solve immediate, practical challenges. For instance, in data-saturated domains, how should we prioritize specific samples to maximize training efficiency? Conversely, in data-scarce domains, which adjacent fields can most effectively augment the primary dataset to bolster performance? Finally, for data-rich target domains, we must identify the specific conditions under which it becomes computationally beneficial to diversify training corpora with samples from distanced disciplines.

Resolving these questions necessitates the construction of quantitative, computationally efficient measures of transfer. The field of deep active learning (DAL) has already developed efficient surrogates to estimate the impact of individual sample inclusion on a target objective.⁹ While these strategies have proven effective for human-in-the-loop data mining and intradomain sample selection, we need to generalize these methodologies to evaluate interdomain knowledge transfer and scientific resource allocation. This includes optimizing compute budgets by identifying which simulations will generate the most informative training data and by developing relevant cross-disciplinary benchmarks.

We can deploy this framework across multiple domains – including fluid dynamics, astrophysics, materials science, and biology – to design connectivity graphs that

map knowledge transfer across disparate datasets. By benchmarking these connectivity maps against established knowledge graphs derived from existing literature, we can uncover latent relationships previously unknown to the scientific community.¹⁰ Ultimately, we will empirically validate our findings by training a polymathic model on the domain combinations showing the highest propensity for transfer.

We lack a rigorous metric for “polymathic intelligence.” While we can measure a model’s accuracy in a specific task, we cannot easily quantify how much “knowledge” is being shared between one discipline and another. Solving this requires the development of new benchmarks that account for cross-domain synergy. We must be able to calculate the “connectivity” of the model’s internal representations, for example, ensuring that an insight in physics materially improves the model’s predictive power in climate science.

We need to dive deep into the following questions: How do large-scale scientific models encode domain-specific heuristics – such as proteomic folding principles, chemical reaction mechanisms, or the governing laws of fluid dynamics – and can these internal representations be rendered both interpretable and transferable? Furthermore, how do scientific foundation models facilitate knowledge migration between related disciplines (for example, from mathematics to physics, chemistry, and biochemistry), and can we isolate the underlying transfer mechanisms? The polymathic hypothesis posits that cross-disciplinary knowledge fundamentally enhances performance on specialized tasks; however, we must still determine how to rigorously quantify and optimize this latent transfer.

One of the most obvious techniques to test knowledge transfer is to compare directly the performance of a model pretrained on domain A and fine-tuned on domain B with a model trained only using data from domain B, based on the same amount of compute. Initial small-scale tests and even some medium-scale tests in fluid dynamics have already shown that pretraining on datasets from different disciplines (specifically, pretraining on biological fluids or oceanography simulations) enhances the capability of the AI fluids model for other contexts, such as aerodynamics or astrophysics. However, the approach of small-scale studies of cross-domain knowledge transfer can be computationally expensive and very often is only a binary comparison, ignoring the effects of including many datasets in training, but it will nonetheless help us build key insights.

Alternatively, we can rely on scaling laws, which illustrate performance improvements relative to increased data and compute, to investigate the connections between datasets from various domains. Recognizing that scaling exponents differ across data types – for instance, Wikipedia data impact language model perplexity more than common crawl data – we can imagine examining how performance in certain domains scales with additional data, allowing us to assess the impact of multidisciplinary data. Additionally, by analyzing scaling-exponent variations

against a heuristic measure of domain distance, such as dataset interconnectedness, we can evaluate the effectiveness of our cross-disciplinary approach and inform future expansion strategies.

Scientific models must be validated not merely by predictive accuracy but by their explanatory power. While black-box architectures suffice for linguistic generation, scientific inquiry demands rigorous interpretability; at a minimum, interpretable results catalyze distal discoveries that remain inaccessible via opaque methodologies. I propose a strategy synthesizing symbolic distillation – the translation of neural computations into closed-form equations – with influence tracing to identify the specific training data governing model predictions. Scientific domains provide a unique ontological advantage: established physical laws serve as a definitive testbed for validating these interpretability claims.

The inherent opacity of these black-box models is a primary challenge in deploying machine-learning systems. Consequently, significant research has focused on elucidating the mechanistic pathways underlying transformer performance, primarily within natural language.¹¹ Foundation models trained on scientific data offer a transformative perspective on this problem because their internal interpretations are empirically verifiable. Since the mathematical structures governing scientific generative processes are often known, they provide a ground truth for verifying interpretations across established scales. Specifically, we can map circuits identified via mechanistic interpretability to documented physical phenomena. The recurrence of identical circuits across disparate problems, alongside interpretable influence functions tracing predictions to granular training examples, serves as a robust metric for cross-domain knowledge transfer.¹² Symbolic regression can be used to distill a neural network into a parsimonious mathematical model.¹³ By employing symbolic regression to iteratively replace neural connections with precise symbolic expressions, substantial segments – or the functional entirety – of a network can be codified into human-readable primitives. Transformers, however, have historically resisted such granular decomposition due to the absence of a compatible formal language. To date, a comprehensive symbolic translation of a full-scale transformer remains unachieved.

Another potential method is retrieval-augmented generation (RAG), a strategy gaining traction for both interpretability and out-of-domain generalization.¹⁴ Rather than relying on a model's opaque parametric memory, RAG architectures query external databases of verified examples to inform predictions. This replaces blind trust with inspectable knowledge; researchers can directly analyze attention maps over retrieved examples to isolate the features driving inference, thereby imbuing scientific models with both interpretability and structural robustness. More recently, methods such as “concept steering” have allowed us to identify concepts in large models (most often LLMs). One of the most prominent examples

of concept steering is Anthropic's use of a relevant set of weights to temporarily deploy an alternate Claude model that pushed conversations toward its concept of the Golden Gate Bridge, even when not relevant ("If you ask this 'Golden Gate Claude' how to spend \$10, it will recommend using it to drive across the Golden Gate Bridge and pay the toll").¹⁵ We have seen early signs of the possibility of locating physical concepts in large scientific models; it remains to be seen whether concepts can be identified in a number of other scientific models.¹⁶

Science is fundamentally multimodal; no singular format predominates and complex phenomena are most profoundly understood through integrated representations. A primary architectural strategy involves engineering universal tokenization schemes (for scientific data, tokenization usually means some kind of compression scheme, and is somewhat different from its linguistic counterparts) for archetypal data structures – such as time series, multidimensional snapshots, and geometric meshes – to align disparate inputs in a unified latent space. Metadata are encoded via language models and appended as discrete token sequences, empowering the foundation model to autonomously calibrate its own preprocessing pipeline. This enables a single generalist system to navigate domains as disparate as heliophysics and developmental biology without sacrificing domain-specific precision. Looking forward, we anticipate that autonomous AI-code synthesis will facilitate the dynamic design of novel tokenization schemes tailored to incoming metadata, enabling the automated curation, compression, and seamless integration of novel data streams into the existing model.

In the scientific domain, datasets lack predefined standardization; every instrument and simulation generates unique output formats and modalities. Consequently, there is no universal preprocessing protocol capable of standardizing these heterogeneous sources. These dual challenges – the absence of universal preprocessing and the massive multimodality of scientific data – imply that current foundation models, which typically operate on a restricted set of well-defined modalities, cannot be readily applied to the vast landscape of scientific inquiry.

We can address these challenges by developing robust tokenization strategies for finite data archetypes, including 1D-3D snapshots, time series, and graph structures. Unlike the nearly infinite variety of modalities, these foundational data archetypes are finite. By creating a specific tokenization strategy for each archetype, we can embed raw observational data from all modalities in a shared high-dimensional representation. To ensure the model correctly interprets this raw tokenized data, it must be provided with sufficient contextual information to perform the equivalent of manual preprocessing autonomously. For this purpose, we integrate comprehensive metadata – describing instrument features and collection parameters – processed through a language model to extract compressed representations as metadata tokens. This approach explicitly incorporates the preprocessing stage into the

model architecture itself, leveraging metadata to align data from disparate instruments and domains within the same latent space. We can validate this methodology through a phased expansion of supported data archetypes. By quantitatively assessing the performance gains achieved through multimodal object representations and benchmarking against specialized models, we will develop a truly universal scientific architecture capable of addressing the full spectrum of archetypes across our target domains.

What fundamental architectural principles enable the seamless integration of linguistic reasoning with comprehensive scientific data and knowledge? We must determine whether to iterate on current agentic paradigms, optimize tool-use protocols, lean into reinforcement learning, or develop something else entirely. Alternatively, while a mixture-of-experts approach offers a path forward, we must investigate if more-natively integrated frameworks can bridge the existing rift between frontier LLMs and scientific foundation models. While there is a large community looking into another paradigm beyond our predominantly transformer- and diffusion-based architectures, there hasn't yet been a breakthrough technology that can outperform current frontier models.¹⁷ Today, the vast disparities between linguistic and scientific models in data scale, parameter counts, and training objectives exacerbate this divide, complicating the creation of a unified framework capable of synthesizing both the linguistic descriptions of science and the raw knowledge encoded in empirical data.

Consider Newton's laws: They are celebrated as elegant equations accompanied by textual explanations, yet their true essence is captured in the heterogeneous observations of a falling apple and the celestial mechanics of orbiting planets. These seemingly disparate data streams are intrinsically linked by a single, parsimonious mathematical set. This raises a critical architectural question: How can we design a system capable of learning across these vast scales and multimodal data types to autonomously derive such aesthetically pleasing and physically accurate equations? While this remains an open challenge, I advocate engineering systems specifically designed for the deep integration of language-based reasoning with expansive scientific datasets. Current model-merging techniques remain insufficient to unify models with radically different architectures. Consequently, achieving true polymathic intelligence requires a paradigm shift, ensuring models can reason fluently across natural language, molecular structures, formal equations, and complex experimental outputs.

The polymathic model is not merely a better predictor. It is a new kind of epistemic tool, one that augments the scientist's imagination, accelerates hypothesis generation, automates data analyses, and reveals hidden commonalities across domains. I imagine that scientists will be assisted by the model in ways that will accelerate scientific discovery. What could scientists accomplish

if some of the most time-consuming work – from collating research materials and digesting materials across different disciplines to even more mundane tasks like curating and cleaning datasets and setting up the codebase with the right environment – became largely automatic? At the same time, the risks are real. Overreliance on automated inference, overgeneralization, and unexamined biases in training data could mislead rather than enlighten. But these are governance problems, not showstoppers. With rigorous benchmarks, open-science principles, and collaboration, these risks can be mitigated.

I believe that the future of scientific discovery will be significantly accelerated by multidisciplinary scientific artificial intelligence – virtual polymaths with expertise spanning every domain – that can be spawned at scale to help address the most challenging scientific questions. Such advances will likely be most prominent in areas that require huge numbers of searches in an immense parameter space, as the power of scale will there be greatest. Scientists can enter the process where it is most critical: examining biases in every aspect of the discovery process, breaking down general misconceptions that may be inherent in the dataset, and coming up with genuinely new out-of-the-box insights and theorems to test.

At its core, my vision centers on a multidisciplinary scientific intelligence that can integrate scientific knowledge spanning all domains within the entirety of the process of discovery: planning, executing, and extracting insights with intentional human intervention. Scientists will be empowered by frontier, multidisciplinary foundation models along with autonomous research and development agents, native access to comprehensive scientific datasets, and high-performance computing resources. The transformative power of this approach lies in its ability to experiment and collaborate across millions of parallelized tasks and overcome the information silos that impair traditional research. This vision is several steps ahead of the current trend of specific foundation model development. A polymathic foundation model will not replace scientists; it will empower them. It offers a scaffolding for interdisciplinary thinking, a bridge across the fragmented landscape of modern research.

To build a generalist AI model for science is more than a technical undertaking. It challenges the very structure of knowledge. Just as polymaths once transformed entire worldviews by connecting what was once thought separate, so too can polymathic models illuminate the hidden unities of nature. I envision a near future in which AI systems assist scientists not just in data analyses but in testing competing theories, extrapolating across scales, and identifying conceptual gaps. To get there, we must invest in shared infrastructure and develop quantitative and scientifically sound benchmarks, novel measures of transfer, and trust. The age of fragmented genius does not need to define AI in science. With intention and collaboration, we can build models that think more like polymaths – and in doing so, extend the reach of human scientific discovery.

ABOUT THE AUTHOR

Shirley Ho is the PI and founder of Polymathic AI, a nonprofit research collaboration of over forty international scientists creating large frontier models in sciences ranging from fluid dynamics to genomics. She is the Group Leader and Senior Research Scientist of the Simons Foundation's Cosmology X Data Science team. She is also faculty in the Department of Physics and in the Center for Data Science at New York University, and previously served as faculty at Princeton University, University of California, Berkeley, and Carnegie Mellon University. She is the recipient of the Carnegie Science Award and NASA Achievement Award, and more recently was selected as a Schmidt Sciences AI2050 Senior Fellow. She has published in *Nature*, *Proceedings of the National Academy of Sciences*, and all the major machine learning conferences.

ENDNOTES

- ¹ John Jumper, Felix Stocker, Pushmeet Kohli, et al., “Highly Accurate Protein Structure Prediction with AlphaFold,” *Nature* 596 (7873) (2021): 583–589.
- ² Siyu He, Yin Li, Yu Feng, Shirley Ho, et al., “Learning to Predict the Cosmological Structure Formation,” *Proceedings of the National Academy of Sciences* 116 (28) (2019): 13825–13832; ChangHoon Hahn, Pablo Lemos, Liam Parker, et al., “Cosmological Constraints from Non-Gaussian and Nonlinear Galaxy Clustering Using the Simbig Inference Framework,” *Nature Astronomy* 8 (11) (2024): 1457–1467; Pablo Lemos, Liam Parker, ChangHoon Hahn, et al., “Field-Level Simulation-Based Inference of Galaxy Clustering With Convolutional Neural Networks,” *Physical Review D* 109 (2024): 083536; Francois Lanusse, Liam Parker, Siavash Golkar, et al., “AstroCLIP: Cross-Modal Pre-Training for Astronomical Foundation Models,” arXiv (2023), <https://doi.org/10.48550/arXiv.2310.03024>; and Ethan Silver, R. Wang, Xiaosheng Huang, et al., “ML-Driven Strong Lens Discoveries: Down to $\theta_E \sim 0.03''$ and $M_{\text{halo}} < 1011 M_\odot$,” arXiv (2025), <https://doi.org/10.48550/arXiv.2507.01943>.
- ³ Liam Parker, Francois Lanusse, Jeff Shen, et al., “Aion-1: Omnimodal Foundation Model for Astronomical Sciences,” arXiv (2025), <https://doi.org/10.48550/arXiv.2510.17960>.
- ⁴ Ryan Roussel, Juan Pablo Gonzalez-Aguilera, Auralee Edelen, et al., “Efficient 6-Dimensional Phase Space Reconstructions from Experimental Measurements Using Generative Machine Learning,” *Physical Review Accelerators and Beams* 27 (2024), <https://doi.org/10.1103/PhysRevAccelBeams.27.094601>.
- ⁵ Diptarka Hait, Jan D. Estrada Pabón, Martin Stöhr, and Todd J. Martínez, “Locating Ab Initio Transition States via Geodesic Construction on Machine-Learned Potential Energy Surfaces,” *Journal of Chemical Theory and Computation* 21 (22) (2025): 11632–11644; and Dario Coscia, Pim de Haan, and Max Welling, “BLIPs: Bayesian Learned Interatomic Potentials,” arXiv (2025), <https://doi.org/10.48550/arXiv.2508.14022>.
- ⁶ Leo Gao, Stella Biderman, Sid Black, et al., “The Pile: An 800GB Dataset of Diverse Text for Language Modeling,” arXiv (2020), <https://doi.org/10.48550/arXiv.2101.00027>; and Alex Wang, Amanpreet Singh, Julian Michael, et al., “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding,” arXiv (2018; rev. 2019), <https://doi.org/10.48550/arXiv.1804.07461>.
- ⁷ The Multimodal Universe Collaboration (Eirini Angeloudi, Jeroen Audenaert, Micah Bowles, et al.), “The Multimodal Universe: Enabling Large-Scale Machine Learning with 100TB of

- Astronomical Scientific Data,” arXiv (2024), <https://doi.org/10.48550/arXiv.2412.02527>; RCSB Protein Data Bank, <https://www.rcsb.org>; Ruben Ohana, Michael McCabe, Lucas Meyer, et al., “The Well: A Large-Scale Collection of Diverse Physics Simulations for Machine Learning,” arXiv (2024; rev. 2025), <https://doi.org/10.48550/arXiv.2412.00568>; and Neil Ashton, Charles Mockett, Marian Fuchs, et al., “DrivAerML: High-Fidelity Computational Fluid Dynamics Dataset for Road-Car External Aerodynamics,” arXiv (2024; rev. 2025), <https://doi.org/10.48550/arXiv.2408.11969>.
- ⁸ For example, see U.S. Department of Energy, “Genesis Mission,” <https://genesis.energy.gov> (accessed March 25, 2026).
- ⁹ Dongyuan Li, Zhen Wang, Yankai Chen, et al., “A Survey on Deep Active Learning: Recent Advances and New Frontiers,” arXiv (2024), <https://doi.org/10.48550/arXiv.2405.00334>; and Tianjiao Wan, Kele Xu, Ting Yu, et al., “A Survey of Deep Active Learning for Foundation Models,” *Intelligent Computing* 2 (2) (2023): 0058.
- ¹⁰ Shilpa Verma, Rajesh Bhatia, Sandeep Harit, and Sanjay Batish, “Scholarly Knowledge Graphs through Structuring Scholarly Communication: A Review,” *Complex & Intelligent Systems* 9 (1) (2023): 1059–1095.
- ¹¹ Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov, “Locating and Editing Factual Associations in GPT,” arXiv (2023), <https://doi.org/10.48550/arXiv.2202.05262>; Eric Todd, Millicent L. Li, Arnab Sen Sharma, et al., “Function Vectors in Large Language Models,” arXiv (2023), <https://doi.org/10.48550/arXiv.2310.15213>; Kevin Wang, Alexandre Variengien, Arthur Conmy, et al., “Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 Small,” arXiv (2022), <https://doi.org/10.48550/arXiv.2211.00593>; Atticus Geiger, Zhengxuan Wu, Christopher Potts, et al., “Finding Alignments Between Interpretable Causal Variables and Distributed Neural Representations,” arXiv (2023; rev. 2024), <https://doi.org/10.48550/arXiv.2303.02536>; and Zhengxuan Wu, Atticus Geiger, Christopher Potts, and Noah D. Goodman, “Interpretability at Scale: Identifying Causal Mechanisms in Alpaca,” arXiv (2023; rev. 2024), <https://doi.org/10.48550/arXiv.2305.08809>.
- ¹² Roger Grosse, Juhan Bae, Cem Anil, et al., “Studying Large Language Model Generalization With Influence Functions,” arXiv (2023), <https://doi.org/10.48550/arXiv.2308.03296>.
- ¹³ John R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection* (MIT Press, 1992); Miles Cranmer, Alvaro Sanchez-Gonzalez, Peter Battaglia, et al., “Discovering Symbolic Models from Deep Learning with Inductive Biases,” in *Advances in Neural Information Processing Systems* 33 (*NeurIPS* 2020), ed. Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, et al. (Curran Associates, Inc., 2020), <https://proceedings.neurips.cc/paper/2020/file/c9f2f917078bd2db12f23c3b413d9cba-Paper.pdf>; and Pablo Lemos, Niall Jeffrey, Miles Cranmer, Peter Battaglia, and Shirley Ho, “Rediscovering Newton’s Gravity and Solar System Properties Using Deep Learning and Inductive Biases,” *Machine Learning: Science and Technology* 4 (2023).
- ¹⁴ Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” arXiv (2020; rev. 2021), <https://doi.org/10.48550/arXiv.2005.11401>; and Lili Yu, Bowen Shi, Ramakanth Pasunuru, et al., “Scaling Autoregressive Multi-Modal Models: Pretraining and Instruction Tuning,” arXiv (2023), <https://doi.org/10.48550/arXiv.2309.02591>.
- ¹⁵ Adly Templeton, Tom Conerly, Jonathan Marcus, et al., “Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet,” Anthropic, Transformer

Circuits Thread, May 21, 2024, <https://transformer-circuits.pub/2024/scaling-mono-semanticity/index.html>.

- ¹⁶ Andy Arditi, Oscar Obeso, Aaquib Syed, et al., “Refusal in Language Models Is Mediated by a Single Direction,” in *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, ed. A. Globerson, L. Mackey, D. Belgrave, et al. (Curran Associates, Inc., 2024).
- ¹⁷ Adrien Bardes, Quentin Garrido, Jean Ponce, et al., “Revisiting Feature Prediction for Learning Visual Representations from Video,” arXiv (2024), <https://doi.org/10.48550/arXiv.2404.08471>.

How Do We Build AI to Push the Frontiers of Scientific Discovery?

Anima Anandkumar

Scientific progress is limited not by a lack of new ideas but by the time and cost involved in physical experimentation. Scientific discovery is a needle in the haystack problem: it does not help if AI gives you a vastly bigger haystack. Without knowing if any of the ideas work, an AI system that designs experiments just increases the effort required, since performing the experiments to validate the ideas is the real bottleneck. In my view, AI's most transformative impact in enabling scientific discoveries lies in reducing the need for such experiments. To get there, we need to build AI models that are able to granularly simulate and understand physics at all scales, rather than just abstractly reason in the textual domain. In this essay, I explore what methods like Neural Operators have already helped achieve, what still needs to be done, and what lies ahead.

Can AI generate truly novel and surprising scientific discoveries and inventions? My research has focused on this challenge for almost a decade. Before my arrival at Caltech, I led AI research at Amazon Web Services, building some of the first AI products on the cloud. Many of my colleagues at Caltech had a healthy dose of skepticism about AI and whether it could have a lasting impact in scientific domains, where data are nowhere as plentiful as images and text on the internet. This challenge of developing AI algorithms for scientific advances became my research focus both at Caltech and during my time at NVIDIA.

Fast forward to this era of large language models, and AI is making significant strides in scientific domains. Language models have achieved gold medal-level performance in major math contests. They have fundamentally changed how many people write software code today. There have also been extensive efforts to use them for scientific discovery. For instance, “AI scientists” have shown some success in specific cases, such as predicting chemical synthesis and drug repurposing.¹ These AI scientists – systems built on language models, sometimes with some additional domain-specific fine-tuning – attempt to generate new ideas and hypotheses based on their textual knowledge. And while such language models may be part of the process of scientific discovery, they are not the complete answer.

Textual models do not address the critical bottleneck of scientific discovery. Scientific progress is not limited by a lack of new ideas but by the cost and time

needed to test them through physical data acquisition and validation in lab experiments. Scientific discovery is a needle in the haystack problem: it does not help if AI gives you a vastly bigger haystack. Without knowing if any of the ideas work through experimental validation, the volume just increases effort. AI's most transformative potential for enabling scientific discoveries therefore lies in reducing the need for such experiments.

To reach this potential, the hypotheses proposed by AI need to have some level of certainty of succeeding in the physical world. This is unlikely to come solely from textual data. For AI to have physical understanding, we need to train it on data from the physical world. There have been many recent success stories of training AI on domain-specific data, with outcomes such as discovering new functional enzymes and predicting coronavirus variants of concern, protein dynamics, and extreme weather events.² One common theme in all these examples is the availability of sufficiently large (curated) datasets. What about areas in which such datasets are unavailable or too difficult to collect? Moreover, even in the above areas, can we go beyond the regime of training data? This is not just a theoretical question. For example, if we are trying to predict extreme weather events in the context of significant climate change, such data are not available.³ In protein folding, many important protein families are not well represented in the current training data.⁴ Further, many bold claims of AI-driven discovery have seen pushback. For instance, the revelation that AI had discovered “2.2 million structures . . . which escaped previous human chemical intuition” was tempered by experiments showing that many of the AI-discovered materials did not function as predicted.⁵

Discovery is about going beyond what is known; it cannot just rely on examples. Standard data-driven AI struggles to extrapolate beyond the training distribution and is typically overconfident and miscalibrated. This implies that purely data-driven AI for scientific discovery has significant limitations, and initial successes may be attributed to either pure luck or being close enough to the training data, making it an interpolation rather than an extrapolation problem.

For AI to make truly novel scientific discoveries, AI models need a broader physical understanding beyond a narrow data distribution. This can come from the integration of a variety of sources and domain constraints – such as the governing laws of physics and symmetries – with data-driven AI.⁶ In addition to improving extrapolation capabilities, the knowledge of physics allows for verification of AI-discoveries, telling us what works in the real world and minimizing or even eliminating the need for physical prototyping. For instance, many physical phenomena are described by partial differential equations (PDE) such as fluid dynamics, plasma evolution, and quantum mechanics. We can train AI models to follow these guardrails of physical laws and domain constraints, making their discoveries more trustworthy.

To build AI that understands the physical world, we need to consider the governing processes that may span from subatomic to planetary scales and involve spatio-temporal evolution of processes across different scales and domains. Interactions between Earth's complex physical processes span more than twelve orders of magnitude in space and time, from micrometers at the molecular scale to thousands of kilometers at the planetary scale.⁷ Analyzing how coronavirus spreads in respiratory aerosol, for example, spans from atomic-scale electronic structure to whole aerosol particle morphology.⁸ For modeling such physical phenomena, we need AI that can capture fine details accurately without being computationally intractable. Standard neural networks such as transformers and convolutional neural networks are ill-equipped to capture multiscale phenomena since they assume fixed-sized inputs and outputs. While such neural networks can learn on discrete distributions such as text or images at a fixed resolution, they are not sufficient to handle multiple scales of interaction in scientific domains. This limits us from using data sources that are available at multiple resolutions and varying levels of fidelity – for instance, weather and climate models run at different resolutions – and limits us from imposing the physics constraints of governing laws at multiple scales.

I led the team of researchers that invented Neural Operators as a universal framework for learning mappings between function spaces.⁹ The relationship between neural networks and Neural Operators mirrors raster versus vector graphics: the former uses a fixed number of pixels and gets blurry when zoomed in, while the latter uses functional representation that retains sharpness even when zoomed in. Neural Operators learn targets that are continuous functions and that can be discretized to any resolution, while achieving super-resolution that goes beyond training data. They can learn from data at multiple scales and incorporate the knowledge of mathematical equations to fill in the finer details when only limited-resolution data are available. Neural Operator-based models can seamlessly assimilate experimental and observational data and adapt to new information. Moreover, since AI models are differentiable, they enable inverse design, supporting exploration and optimization under diverse design and manufacturing constraints.

Neural Operators have already achieved success in many scientific domains. I am proud to have led an interdisciplinary team from NVIDIA, Caltech, Berkeley, and other universities that trained FourCastNet, the first high-resolution, fully AI-based weather model.¹⁰ FourCastNet has competitive accuracy while being tens of thousands of times faster than traditional weather models. This surprised many weather scientists, given the general consensus that AI was still in its infancy and that a number of fundamental breakthroughs would be needed before deep learning could be competitive with traditional weather models, which could take years or even decades.¹¹ FourCastNet led Huawei and Google DeepMind,

among many others, to pursue AI-weather models and for weather agencies such as the European Centre for Medium-Range Weather Forecasts to host AI-weather models alongside traditional forecasts, allowing the community to compare the effectiveness of AI models in action.¹²

FourCastNet also correctly predicted many extreme weather events, including Hurricane Lee and Beryl, several days before traditional weather models.¹³ This was surprising since there are not many instances of hurricanes and other extreme events in the training data. How can AI weather models do so well even with rare events? One possible explanation is that hurricanes carry the distinctive physical signature of warm ocean air rising into the storm, which could be easily learned with only a few samples.¹⁴ It is noteworthy that the overall training of FourCastNet used only about fifty thousand samples, which is minuscule compared with the web-scale training of language models. In other domains, we have seen similar instances of physical phenomena that were thought to be prohibitively complex but do not actually require an inordinate amount of training data.¹⁵ But access to *high-quality* curated data is another important common feature among these success stories. In the case of weather, the open dataset ERA5 consists of roughly forty years of reanalysis weather data, which combine raw satellite observations with physics. In this sense, the AI-weather models are already physics-informed, and thus have better generalization abilities, which makes them practically useful.

The claims of AI improving our ability to predict extreme weather events must be carefully tested, and a few success stories with recent hurricanes do not suffice to make a definitive scientific claim. How can models prove themselves? Extreme weather is a chaotic system. Popularly known as the “butterfly effect,” small changes in initial conditions of a system can lead to rapid divergence in trajectories and radically different outcomes. This means that an already-chaotic system like extreme weather is fundamentally unpredictable beyond a certain time horizon using deterministic forecasts.¹⁶ Prediction models can overcome this unpredictability either by incorporating probabilistic outputs, as with diffusion models, or by considering statistical ensembles over initial conditions. The latter is more lightweight, and due to the speed of deterministic AI-weather models such as FourCastNet, we can support large ensembles that are out of reach for traditional weather models. Further, since we can test for physical validity of each ensemble member and filter out the nonphysical ones, we can better align with ground-truth physics. We train on such ensembles to accurately capture the tail event probability in the latest iteration of training for FourCastNet (version three).¹⁷ This achieves calibrated probabilistic performance on extreme weather with speed and accuracy. Using physically realistic large-ensemble weather forecasts up to sixty days ahead, FourCastNet3 surpasses traditional and AI-weather models in medium-range skill, speed, and stability, while providing trustworthy predictions. This has immense promise for early-warning systems for catastrophic weather events worldwide.

For an AI model to make a broader impact, it needs to have a broader physical understanding, as opposed to being just a narrow surrogate limited to its training regime. Surprisingly, FourCastNet functions not only as a weather model for the short to medium term but is also impactful for subseasonal and seasonal modeling and even long-term climate modeling.¹⁸ Given that the pretraining of FourCastNet involves one-step prediction, typically over six hours, with some additional multistep fine-tuning over a small number of steps in autoregressive rollouts, it is surprising that FourCastNet is able to extrapolate to much longer timescales of several months to years. This capability does not exist in other AI-weather models, which experience blowups over long rollouts. This long-term stability can be attributed to the Spherical Fourier Neural Operator architecture employed in FourCastNet.¹⁹ This design incorporates multiscale learning into an equivariant neural architecture that enforces spherical geometry (is equivariant to rotations on the sphere). In contrast, other AI-weather models use standard neural networks, such as transformers or graph neural networks, which assume a rectangular domain, leading to buildup of distortions near the poles and eventually to catastrophic blowups. Geometry-informed AI has been more robust for extrapolation in other areas as well, such as in molecular and protein modeling.²⁰ This demonstrates the importance of AI being informed by domain geometry and symmetries, and incorporating them into the neural architectures to achieve broad generalization.

Careful design of FourCastNet has given it broad capabilities, earning it the label of foundation model. But the extrapolation capabilities of such models are not boundless. Some AI weather models are unable to forecast “gray swans,” the extreme weather events that are so rare they do not exist in the training set.²¹ Yet such events are possibilities in the context of climate change, despite the absence of observational data.²² Similarly, in other areas such as protein folding, the predictions degrade with proteins outside the training set.²³ As discussed earlier, true extrapolation may not occur using purely data-driven methods. We need physics-informed AI to push those boundaries further.

While the machine-learning community has focused on data-driven approaches, the applied math and computational sciences community has prioritized modeling physical phenomena through partial differential equations and other mathematical equations, and applying numerical methods for solving them. Compared with machine-learning methods that can be considered top-down, mathematical modeling is bottom-up and provides a mechanistic explanation of the observed phenomenon. PDEs are popular in mathematical modeling because they relate across space and time and can model diverse phenomena such as fluid dynamics, material deformation, and quantum chemistry. However, numerical solvers for PDEs are computationally expensive since they require a fine grid to capture the multiscale nature of dynamics of many phenom-

ena. Thus, many real-world phenomena are intractable to simulate accurately even though their governing laws can be written down as succinct PDEs. For instance, Schrodinger's equation models the behavior of all matter, but to solve it exactly using current numerical methods for even a small molecule of about one hundred electrons would take more than the age of the universe. As another example, today's global climate models are unable to resolve clouds and other fine-scale phenomena, which are the biggest source of uncertainty for climate change predictions. Resolving them with current numerical methods is estimated to require several orders of magnitude more computing than is available today.²⁴

Given such impossible computational requirements for exact numerical solutions of PDEs, many approximations have been developed, including density functional theory (DFT) for quantum chemistry and large eddy simulations (LES) for fluid dynamics.²⁵ However, deep domain expertise may be needed to choose the appropriate level of approximation for a specific problem, and such approximations still may not be optimal. In many scenarios, even the approximate numerical solutions are still too expensive to be practical, such as for DFT on protein dynamics, among other simplifications that may not be well principled. Can AI come up with novel approximations that are both significantly cheaper and accurate? The revolution in computer vision and image processing has already shown this is possible: while methods prior to deep learning that used hand-engineered features struggled to reach good accuracy, feature learning with neural networks has been far more successful. Similarly, Neural Operators can learn faster solvers for PDEs compared with human-engineered numerical methods. Additionally, they can incorporate experimental and observed data to reduce modeling errors. We already see this with the Neural Operator-based weather model described above, which is tens of thousands of times faster than physics-based forecasts that solve PDEs, and which also surpasses them in accuracy since it can overcome modeling errors.

An alternative approach to data-driven methods for solving PDEs is Physics-Informed Neural Networks (PINNs), which attempt to solve PDEs in a data-free manner by imposing the PDE residual as the loss function.²⁶ PINNs can be viewed as test-time optimization, and the neural network serves as the *ansatz* for optimization. However, for many time-varying practical problems, the optimization landscape is often intractable, and even when they converge, PINNs are typically not faster than numerical solvers.²⁷ In contrast, we can train Neural Operators to learn the solution operator of PDEs via supervised learning by generating solutions to PDEs using existing numerical simulations. Such a data-driven approach overcomes the optimization challenges of PINNs. However, it is often expensive to generate sufficient high-quality data to train AI. A more-nuanced approach could combine simulation data from a variety of numerical solvers at different levels of approximation as well as incorporate the knowledge of physics, either as PDE residual loss function or through symmetry and conservation laws. Neural Operators are ideal

for this, since the data from numerical solvers at different levels of approximation are typically at different resolutions; for example, the direct-numerical simulation in fluid dynamics is on a finer grid than is large-eddy simulation, and the PDE residual can be imposed as a loss function at different resolutions.

Neural Operators present a principled AI approach for solving PDEs and modeling processes on a continuum, since they are not restricted to one resolution or scale.²⁸ They learn operators that are mappings between function spaces. They are discretization-invariant, meaning they work on any discretization of inputs without retraining. Neural Operators are universal approximators in function spaces: they can fit any continuous operator, such as solutions of PDEs at all resolutions.²⁹ This grid-free approach of Neural Operators allows us to seamlessly train on data at multiple resolutions. We can also impose physics losses, such as PDE residuals, at a higher resolution than the training data. Such a physics-informed Neural Operator approach can potentially surpass the accuracy of the training data.³⁰ Neural Operators can therefore holistically combine data and physics at multiple scales, harmonizing the bottom-up physics-based modeling and top-down data-driven learning.

The development of Neural Operators comes from first principles. Let us start with the simple setting of linear PDEs, for which the solution operator can be expressed as a linear integral operator with a given kernel, also known as Green's function. These kernel integrations can be learned using graphs, Fourier transforms, or attention mechanisms. We can then extend this idea for learning general nonlinear operators by composing learnable linear integral operators with nonlinear activations. This general framework of Neural Operators can universally approximate any nonlinear continuous operator, including solutions to nonlinear PDEs. Many popular neural-network architectures can be converted into Neural Operators that work at any resolution – such as the Fourier Neural Operator, Graph Neural Operator, and Transformer Neural Operator – yielding a rich family of models that are expressive and carry good inductive biases relevant to scientific domains.³¹

In many areas, Neural Operators have already replaced numerical solvers at rates that are four to six orders of magnitude faster. In addition to weather forecasting, this approach has been transformative in other areas such as modeling plasma evolution in nuclear fusion for detecting disruptions and modeling carbon dioxide storage in reservoirs for climate change mitigation, as well as detecting deviations from assumed physics, such as tipping points in climate change that present significant deviation from current conditions.³² Further, Neural Operators enable inverse design since they are differentiable and can both provide guidance for iterative design improvements and generate new designs from scratch. For instance, Neural Operators optimally designed a novel medical catheter that reduces bacterial contamination by two orders of magnitude, addressing one of the most common

sources of health care–related infections.³³ Neural Operators accurately simulated the bacterial density in fluid flow and designed shapes within the catheter to prevent bacteria from swimming upstream into the human body. The AI-designed catheter required only one set of physical experiments for validation and removed the need for the laborious trial-and-error loop typical of design optimization. In inverse lithography, Neural Operators produced better mask quality while being six times faster through progressive self-training, which is not feasible with a traditional solver.³⁴ In lung ultrasound imaging, Neural Operators can work on raw radio frequency signals directly and be physics-informed, yielding much better accuracy in lung aeration levels, as opposed to the current practice of subjective B-mode evaluation.³⁵ Diffusion Neural Operators can learn better priors from data when solving inverse problems.³⁶ Neural Operators are also helpful in deriving efficient control policies through reinforcement learning on hard problems such as drag reduction in turbulent systems.³⁷ Thus, the approach of using physically valid AI-generated designs, discoveries, and control policies using Neural Operators has been transformative in many areas.

Neural Operators also offer a foundation for novel scientific discovery involving multiscale phenomena. Combining this with symbolic reasoning can further boost capabilities for extrapolation. For instance, proving theorems about phenomena such as fluid dynamics would entail symbolic reasoning. PINNs have helped make progress on the famous Millennium Prize problem in fluid dynamics, but the symbolic analysis has so far been done by humans. Can we build AI that can do both?³⁸ There has been exciting progress in theorem-proving using AI, with AI systems turning in gold-level performances at the International Math Olympiad contest for high school students. Several of the winning AI systems made use of verifiers like Lean, a formal proof system that can check every proof step proposed by language models. This ensures hallucination-free reasoning, which is essential for AI to prove long theorems: in the absence of verifiers, errors can build up catastrophically over long reasoning chains. I led a team of researchers in the development and release of the first open-source framework combining language models with Lean – LeanDojo – which has ushered in this revolution in math AI.³⁹ This progress in formal reasoning is complementary to physical modeling with Neural Operators, as it allows us to refine PDEs and their numerical approximations both for data generation and physics losses to train Physics-Informed Neural Operators, while guaranteeing their correctness.

Neural Operators can be thought of as an end-to-end approach to using AI in scientific research. Conversely, more-traditional approaches have retained numerical workflows with traditional solvers and augmented them with AI, typically as an additive correction term. Popular in domains such as fluid dynamics, plasma physics, and climate modeling, this is known as “closure

modeling.”⁴⁰ More generally, in areas like molecular modeling, it is known as coarse graining and renormalization. This process, which predates the use of machine learning, corrects approximate coarse-grid numerical solvers using closures built from physics, such as large eddy simulation for fluid dynamics.⁴¹ Thus, there are two paradigms for physics-informed learning: replace with AI or augment with AI. The former incorporates physics constraints into end-to-end learning while the latter augments existing numerical solvers with learned closures. Is there a clear winner?

In terms of computational efficiency, end-to-end learning with Neural Operators offers significant speed gains over traditional solvers (four to six orders of magnitude), along with differentiability for inverse problems. On the other hand, with closure modeling, we typically do not get the full benefits of faster models and, often, the orchestration of numerical solvers with machine learning leads to slowdowns rather than speedups. Closure-model systems are also typically non-differentiable and therefore not suitable for efficiently solving inverse problems through gradient descent.

With regard to training data requirements, one would expect closure models to be more data efficient, since they would not need to learn from scratch due to augmentation with coarse-grid numerical solvers. We proved a surprising result to the contrary: learning-based closure models fare poorly in both data efficiency and training stability compared with end-to-end learning. Because closure models only have access to training data on a coarse grid, the learning closures are an ill-posed problem. Further, corrections made by the closure model are only allowed additively on the same coarse grid as the numerical solver, which limits expressivity and does not allow nonlinear mixing between scales. We proved that Neural Operators not only circumvent the limitations of standard closure models but also achieve statistically optimal estimation of long-term statistics of chaotic systems.⁴² To achieve data efficiency, we progressively trained the Neural Operator at multiple resolutions: first with coarse-grid solver data, which is cheaper to obtain, then by fine-tuning with a small amount of fine-grid simulations and physics constraints. Viewed under this lens, the Neural Operator approach can be seen as a generalization of the closure model, removing its limitations, in which coarse graining is possible through nonlinear mixing across scales and learning can happen progressively using multiresolution, multifidelity training samples and physics knowledge.

The rapid advances in AI occurring today *can* enable novel discoveries, but not without strong algorithmic foundations. Discovery is a journey into the unknown and cannot rely solely on AI trained on known samples; infusing physics into AI is essential. Since our world has physical interactions at multiple scales, we need AI that can traverse resolutions and scales. Neural Operators offer a principled foundation for extending standard neural networks to function spaces to enable multiscale physics-informed learning. A foundation model for physics

built on Neural Operators will be able to solve multiscale, multiphysics problems in a variety of domains. Similar to how complex sentences are formed by the arrangement of and interactions between words and phrases, the real-world environments in which science and engineering researchers work are made up of combinations of different physical systems. A foundation model with physical understanding would be able to learn such a compositionality and attain emergent capabilities to simulate complex physical phenomena and generate novel designs previously out of reach. The development and use of Neural Operators provide a strong foundation for building AI systems with broad physical understanding across scales and domains, equipping us to tackle the full complexity of the real world for scientific discovery.

ABOUT THE AUTHOR

Anima Anandkumar is the Bren Professor of Computing and Mathematical Sciences at the California Institute of Technology. She previously served as Senior Director of AI Research at NVIDIA and Principal Scientist at Amazon Web Services. She has recently published in such journals as *Nature*, *Science Advances*, *Nature Reviews Physics*, *Science Robotics*, and *Journal of Machine Learning Research*.

ENDNOTES

- ¹ Zhiling Zheng, Ali H. Alawadhi, Saumil Chheda, et al., “Shaping the Water-Harvesting Behavior of Metal-Organic Frameworks Aided by Fine-Tuned GPT Models,” *Journal of the American Chemical Society* 145 (51) (2023), <https://pubs.acs.org/doi/10.1021/jacs.3c12086>; and Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, et al., “Robin: A Multi-Agent System for Automating Scientific Discovery,” arXiv (2025), <https://arxiv.org/abs/2505.13400>.
- ² Théophile Lambert, Amin Tavakoli, Gautham Dharuman, et al., “Sequence-Based Generative AI-Guided Design of Versatile Tryptophan Synthases,” bioRxiv (2025), <https://doi.org/10.1101/2025.08.30.673177>; Maxim Zvyagin, Alexander Brace, Kyle Hippe, et al., “GenSLMs: Genome-Scale Language Models Reveal SARS-CoV-2 Evolutionary Dynamics,” *The International Journal of High Performance Computing Applications* 37 (6) (2023): 683–705, <https://doi.org/10.1177/10943420231201154>; Zhuoran Qiao, Weili Nie, Arash Vahdat, et al., “State-Specific Protein-Ligand Complex Structure Prediction With a Multiscale Deep Generative Model,” *Nature Machine Intelligence* 6 (2) (2024): 195–208; and Thorsten Kurth, Shashank Subramanian, Peter Harrington, et al., “FourCastNet: Accelerating Global High-Resolution Weather Forecasting Using Adaptive Fourier Neural Operators,” in *PASC ’23: Proceedings of the Platform for Advanced Scientific Computing Conference* (Association for Computing Machinery, 2023), 1–11, <https://doi.org/10.1145/3592979.3593412>.

- ³ Y. Qiang Sun, Pedram Hassanzadeh, Mohsen Zand, et al., “Can AI Weather Models Predict Out-of-Distribution Gray Swan Tropical Cyclones?” *Proceedings of the National Academy of Sciences* 122 (21) (2025): e2420914122, <https://doi.org/10.1073/pnas.2420914122>.
- ⁴ Christina Bergonzo and Alexander Grishaev, “Critical Assessment of RNA and DNA Structure Predictions via Artificial Intelligence: The Imitation Game,” *Journal of Chemical Information and Modeling* 65 (7) (2025): 3544–3554.
- ⁵ Anthony K. Cheetham and Ram Seshadri, “Artificial Intelligence Driving Materials Discovery? Perspective on the Article: Scaling Deep Learning for Materials Discovery,” *Chemistry of Materials* 36 (8) (2024): 3490–3495; and Mark Peplow, “AI Is Dreaming Up Millions of New Materials. Are They Any Good?” *Nature* 646 (8083) (2025): 22–25, <https://doi.org/10.1038/d41586-025-03147-9>.
- ⁶ Maziar Raissi, Paris Perdikaris, and George E. Karniadakis, “Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations,” *Journal of Computational Physics* 378 (2019): 686–707; and Zongyi Li, Hongkai Zheng, Nikola Kovachki, et al., “Physics-Informed Neural Operator for Learning Partial Differential Equations,” *ACM/IMS Journal of Data Science* 1 (3) (2024): 1–27, <https://doi.org/10.1145/3648506>.
- ⁷ Kurth, Subramanian, Harrington, et al., “FourCastNet: Accelerating Global High-Resolution Weather Forecasting Using Adaptive Fourier Neural Operators.”
- ⁸ Abigail Dommer, Lorenzo Casalino, Fiona Kearns, et al., “#COVIDisAirborne: AI-Enabled Multiscale Computational Microscopy of Delta SARS-CoV-2 in a Respiratory Aerosol,” *The International Journal of High Performance Computing Applications* 37 (1) (2022), <https://doi.org/10.1177/10943420221128233>.
- ⁹ Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, et al., “Fourier Neural Operator for Parametric Partial Differential Equations,” paper prepared for the International Conference on Learning Representations (ICLR 2021), January 12, 2021, <https://openreview.net/forum?id=c8P9NQVtmnO>; and Kamyar Azizzadenesheli, Nikola Kovachki, Zongyi Li, et al., “Neural Operators for Accelerating Scientific Simulations and Design,” *Nature Reviews Physics* 6 (2024): 320–328.
- ¹⁰ Kurth, Subramanian, Harrington, et al., “FourCastNet: Accelerating Global High-Resolution Weather Forecasting Using Adaptive Fourier Neural Operators.”
- ¹¹ Martin G. Schultz, Clara Betancouty, Bing Gong, et al., “Can Deep Learning Beat Numerical Weather Prediction?” *Philosophical Transactions of the Royal Society A* 379 (2194) (2021), <https://doi.org/10.1098/rsta.2020.0097>.
- ¹² Matthew Chantry, Zied Ben Bouallegue, Linus Magnusson, et al., “The Rise of Machine Learning in Weather Forecasting,” European Centre for Medium-Range Weather Forecasts, June 20, 2023, <https://www.ecmwf.int/en/about/media-centre/science-blog/2023/rise-machine-learning-weather-forecasting>.
- ¹³ Gregory Barber, “AI Hurricane Predictions Are Storming the World of Weather Forecasting,” *Wired*, September 27, 2023, <https://www.wired.com/story/ai-hurricane-predictions-are-storming-the-world-of-weather-forecasting>.
- ¹⁴ Kerry A. Emanuel, “The Theory of Hurricanes,” *Annual Review of Fluid Mechanics* 23 (1) (1991): 179–196.
- ¹⁵ Qiao, Nie, Vahdat, et al., “State-Specific Protein-Ligand Complex Structure Prediction With a Multiscale Deep Generative Model.”

- ¹⁶ Julia Slingo and Tim Palmer, “Uncertainty in Weather and Climate Prediction,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 369 (1956) (2011): 4751–4767.
- ¹⁷ Boris Bonev, Thorsten Kurth, Ankur Mahesh, et al., “FourCastNet 3: A Geometric Approach to Probabilistic Machine-Learning Weather Forecasting at Scale,” arXiv (2025), <https://arxiv.org/abs/2507.12144>.
- ¹⁸ Oliver Watt-Meyer, Brian Henn, Jeremy McGibbon, et al., “ACE2: Accurately Learning Subseasonal to Decadal Atmospheric Variability and Forced Responses,” *npj Climate and Atmospheric Science* 8 (1) (2025): 205; and Haiwen Guan, Troy Arcomano, Ashesh Chattopadhyay, and Romit Maulik, “LUCIE-3D: A Three-Dimensional Climate Emulator for Forced Responses,” arXiv (2025), <https://arxiv.org/abs/2509.02061>.
- ¹⁹ Boris Bonev, Thorsten Kurth, Christian Hundt, et al., “Spherical Fourier Neural Operators: Learning Stable Dynamics on the Sphere,” in *Proceedings of the 40th International Conference on Machine Learning* (PMLR, 2023), 2806–2823.
- ²⁰ Qiao, Nie, Vahdat, et al., “State-Specific Protein-Ligand Complex Structure Prediction With a Multiscale Deep Generative Model.”
- ²¹ Sun, Hassanzadeh, Zand, et al., “Can AI Weather Models Predict Out-of-Distribution Gray Swan Tropical Cyclones?”
- ²² Jacob B. Landsberg and Elizabeth A. Barnes, “Forecasting the Future with Yesterday’s Climate: Temperature Bias in AI Weather and Climate Models,” arXiv (2025), <https://www.arxiv.org/abs/2509.22359>.
- ²³ Bergonzo and Grishaev, “Critical Assessment of RNA and DNA Structure Predictions via Artificial Intelligence.”
- ²⁴ Tapio Schneider, João Teixeira, Christopher S. Bretherton, et al., “Climate Goals and Computing the Future of Clouds,” *Nature Climate Change* 7 (1) (2017): 3–5.
- ²⁵ Frank Jensen, *Introduction to Computational Chemistry* (John Wiley & Sons, 2017).
- ²⁶ Raissi, Perdikaris, and Karniadakis, “Physics-Informed Neural Networks.”
- ²⁷ Aditi Krishnapriyan, Amir Gholami, Shandian Zhe, et al., “Characterizing Possible Failure Modes in Physics-Informed Neural Networks,” *Advances in Neural Information Processing Systems* 34 (*NeurIPS 2021*), ed. Marc’Aurelio Ranzato, Alina Beygelzimer, Yann Dauphin, et al. (Curran Associates, Inc., 2021), 26548–26560; and Sifan Wang, Shyam Sankaran, Xiantao Fan, et al., “Simulating Three-Dimensional Turbulence with Physics-Informed Neural Networks,” arXiv (2025), <https://arxiv.org/abs/2507.08972>.
- ²⁸ Li, Kovachki, Azizzadenesheli, et al., “Fourier Neural Operator for Parametric Partial Differential Equations”; and Julius Berner, Miguel Liu-Schiaffini, Jean Kossaifi, et al., “Principled Approaches for Extending Neural Architectures to Function Spaces for Operator Learning,” arXiv (2025), <https://doi.org/10.48550/arXiv.2506.10973>.
- ²⁹ Nikola Kovachki, Zongyi Li, Burigede Liu, et al., “Neural Operator: Learning Maps between Function Spaces with Applications to PDEs,” *Journal of Machine Learning Research* 24 (89) (2023): 1–97.
- ³⁰ Zongyi Li, Hongkai Zheng, Nikola Kovachki, et al., “Physics-Informed Neural Operator for Learning Partial Differential Equations,” *ACM/IMS Journal of Data Science* 1 (3) (2024): 1–27, <https://doi.org/10.1145/3648506>.

- ³¹ Berner, Liu-Schiaffini, Kossai, et al., “Principled Approaches for Extending Neural Architectures to Function Spaces for Operator Learning.”
- ³² Vignesh Gopakumar, Stanislas Pamela, Lorenzo Zanisi, et al., “Plasma Surrogate Modelling Using Fourier Neural Operators,” *Nuclear Fusion* 64 (5) (2024), <https://doi.org/10.1088/1741-4326/ad313a>; Gege Wen, Zongyi Li, Qirui Long, et al., “Real-Time High-Resolution CO₂ Geological Storage Prediction Using Nested Fourier Neural Operators,” *Energy & Environmental Science* 16 (4) (2023): 1732–1741; and Miguel Liu-Schiaffini, Clare E. Singer, Nikola Kovachki, et al., “Tipping Point Forecasting in Non-Stationary Dynamics on Function Spaces,” arXiv (2023), <https://arxiv.org/abs/2308.08794>.
- ³³ Tingtao Zhou, Xuan Wan, Daniel Zhengyu Huang, et al., “AI-Aided Geometric Design of Anti-Infection Catheters,” *Science Advances* 10 (1) (2024), <https://doi.org/10.1126/sciadv.adj1741>.
- ³⁴ Haoyu Yang, Zongyi Li, Kumara Sastry, et al., “Large Scale Mask Optimization via Convolutional Fourier Neural Operator and Litho-Guided Self Training,” arXiv (2022), <https://doi.org/10.48550/arXiv.2207.04056>.
- ³⁵ Jiayun Wang, Oleksii Ostras, Masashi Sode, et al., “Ultrasound Lung Aeration Map via Physics-Aware Neural Operators,” arXiv (2025), <https://doi.org/10.48550/arXiv.2501.01157>.
- ³⁶ Jiachen Yao, Abbas Mammadov, Julius Berner, et al., “Guided Diffusion Sampling on Function Spaces with Applications to PDEs,” arXiv (2025), <https://doi.org/10.48550/arXiv.2505.17004>.
- ³⁷ Zelin Zhao, Zongyi Li, Kimia Hassibi, et al., “Physics-Informed Neural-Operator Predictive Control for Drag Reduction in Turbulent Flows,” arXiv (2025), <https://doi.org/10.48550/arXiv.2510.03360>.
- ³⁸ Yixuan Wang, Ziming Liu, Zongyi Li, et al., “High Precision PINNs in Unbounded Domains: Application to Singularity Formulation in PDEs,” arXiv (2025), <https://arxiv.org/abs/2506.19243>.
- ³⁹ Kaiyu Yang, Aidan M. Swope, Alex Gu, et al., “LeanDojo: Theorem Proving with Retrieval-Augmented Language Models,” in *NIPS ’23: Proceedings of the 37th International Conference on Neural Information Processing Systems* (Association for Computing Machinery, 2023).
- ⁴⁰ Dmitrii Kochkov, Janni Yuval, Ian Langmore, et al., “Neural General Circulation Models for Weather and Climate,” *Nature* 632 (8027) (2024): 1060–1066.
- ⁴¹ Chuwei Wang, Julius Berner, Zongyi Li, et al., “Coarse Graining with Neural Operators for Simulating Chaotic Systems,” arXiv (2025), <https://doi.org/10.48550/arXiv.2408.05177>.
- ⁴² Ibid.

Toward a Science of Intelligence: Unifying Physics, Neuroscience & AI

Surya Ganguli

Artificial intelligence stands poised to transform our society, yet we hardly understand how it works. A synthesis of physics, neuroscience, and AI can fulfill an urgent need: to build a new, unified science of intelligence that explains and improves how intelligence emerges across both artificial and biological neural networks. I discuss four ways this synthesis has begun to and will continue to unfold. First, powerful analytic tools from the physics of complex systems will provide insight into how large neural networks learn and compute. Second, neuroscience will provide clues into bridging the many orders of magnitude advantages that biological intelligence retains over AI. Third, we can go beyond evolution to instantiate neural algorithms in quantum hardware, leading to new devices through AI-led codesign of physics and computation. Fourth, we can meld minds and machines by building digital twins of the brain, yielding insights into not only intelligence but also consciousness and the sense of self through causal modeling and control. Overall, AI will shed light on the nature of our physical and mental realities, raising profound questions about the role of human understanding in the age of AI.

The first glimmers of human-like intelligence appeared in Africa a few million years ago, culminating in the brain of our species *Homo sapiens* about three hundred thousand years ago. Our ancient ancestors no doubt peered both out into the night sky to contemplate the nature of physical reality and inward to ponder the nature of their own mental reality.

In the last century, we have developed a deep understanding of physical reality through precise mathematical laws governing the behavior of space, time, matter, and energy. But we are only beginning to understand our mental reality. How does human intelligence and consciousness emerge from one hundred billion neurons connected by one hundred trillion synapses? Modern neuroscience has laid strong foundations for attacking this grand question. But when it comes to our own mental capabilities, we seek not only to understand but also to recreate these capabilities in machines, sometimes fashioned after our own image. In essence, humans, as products of evolution, yearn to play the role of creator. This yearning permeates literature, ranging from Mary Shelley's *Frankenstein* to Isaac Asimov's *I, Robot*.

In the last decade, artificial intelligence has made striking progress in realizing this yearning, making highly intelligent systems capable of vision, language, reasoning, imagination, and more. These capabilities emerged in an incredibly nonintuitive and *implicit* manner from scaling up large neural networks, fit to large amounts of data, on large compute clusters. Such AI systems stand poised to transform our economy, society, and the very nature of scientific research itself. However, quite alarmingly, such AI systems possess large and mysterious surfaces of *both* capabilities *and* fragilities, and we barely have any scientific understanding of either. Thus, there is a fierce urgency to develop a deeper scientific understanding of AI to ensure the development of trustworthy, robust AI systems aligned with human needs.

There are immense opportunities for advancing our scientific understanding of intelligence, across both brains and machines, through an interdisciplinary synthesis of physics, neuroscience, and AI. First, physics has developed powerful theoretical tools for the analysis of complex systems consisting of many billions of interacting particles with intricate, dynamic laws that influence their collective behavior and give rise to surprising “emergent” properties: new characteristics that arise from the interactions but are not present in the individual particles themselves. Similarly, learning in neural networks involves many billions of interacting weights, with complex learning dynamics driven by data, giving rise to surprising emergent capabilities. Thus, ideas from physics have and will continue to have a powerful impact on AI. Second, neuroscience and the allied fields of psychology and cognitive science reveal multiple orders of magnitude advantages that human intelligence still retains over AI, especially in terms of data efficiency, energy efficiency, and robustness. Neuroscience in this way provides both beacons and clues toward improving AI. Third, one can combine neuroscience and physics to go beyond evolution, to instantiate the neural algorithms for computation discovered by evolution but through quantum hardware that evolution could not discover. This yields new types of quantum neuromorphic technologies inspired by the synthesis of physics and neuroscience. Fourth, AI can advance our understanding of biological intelligence and consciousness by turning back toward us and employing bidirectional feedback loops that meld brains and machines. Thus, while human intelligence gave birth to artificial intelligence, our child will become our teacher and will help us peer into the nature of our own mental reality as well as physical reality itself.

The rest of the twenty-first century will be an exciting intellectual adventure, in which physics, neuroscience, and AI work hand in hand to teach each other and us about the science of intelligence.

Physics has long shaped the development of artificial intelligence. While so-called fundamental physics historically strived to understand nature at the smallest subatomic scales, condensed matter physicist Phillip Anderson argued in his influential 1972 essay “More Is Different” that an equally fundamental

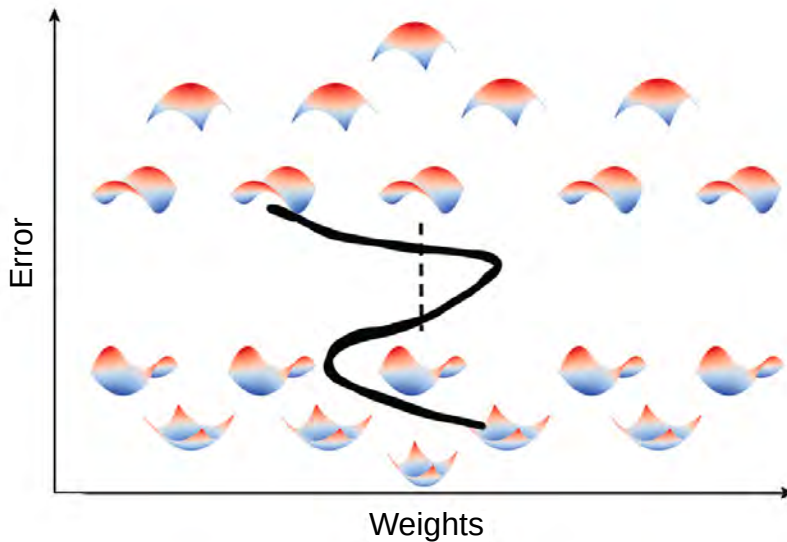
problem involves understanding how surprising and nonintuitive properties at a larger scale *emerge* from the interactions of many simple elements at a smaller scale.¹ For example, how does the “wetness” of water emerge from the interaction of many H_2O molecules? Why do metals conduct? How does glass achieve its simultaneous rigidity and fragility? The field of condensed matter physics arose to develop powerful theoretical tools to analyze complex matter consisting of many interacting atoms.

However, the notion of emergence extends beyond physics. For example, how does cellular life emerge from the interactions of many nonliving molecules? How does intelligence emerge from the interaction of many nonintelligent neurons? Bolstered by their success in condensed matter physics, many physicists turned to biology and neuroscience. John Hopfield was one such physicist: In 1982, he developed a breakthrough model of associative memory through an analogy to magnetic spin systems in physics.² This model explained how networks of neurons could store new facts and recall facts given partial information. In turn, the Hopfield model led to the Boltzmann machine and deep-belief networks, both precursors to modern deep learning.³ For this arc of discovery, John Hopfield and Geoffrey Hinton won the 2024 Nobel Prize in Physics. And this arc continues: the modern transformer architecture is closely related to a modified Hopfield network.⁴

Another influential line of work in the theoretical physics of complex systems is that of Giorgio Parisi, who won part of the 2021 Nobel Prize in Physics. Parisi and colleagues showed how to analyze the geometry of high-dimensional energy landscapes over many particles and describe how the particles move in this energy landscape geometry.⁵ We are used to thinking about functions over two-dimensional space (for instance, the height of an actual hilly landscape on the surface of the earth). But the energy landscape of a physical system is a function of the position of many billions of particles. Similarly, the loss landscape of a neural network is a function of the values of many billions of parameters. Thus, there can be commonalities in how cooled physical systems lower their energy as particle positions change and in how neural networks learn to lower their loss as their weights change. Indeed, some features of the geometry of and dynamics within high-dimensional landscapes can exhibit universal properties. For example, why doesn't neural network learning get stuck in local minima with high error, like a ball rolling down a rugged hilly landscape might get stuck in a high valley? The key idea is that in high but not low dimensions, such high-energy minima are exponentially rare. Intuitively, suppose you are at an energy or loss minimum in a billion-dimensional space. What are the chances that motion in *all* one billion directions will go up? The answer is extremely unlikely, unless you are *already* near the bottom of the landscape. Local minima therefore typically only occur in a range of energies or losses close to that of the global minimum. Instead, saddle points proliferate, which always provide a way down (Figure 1). This intuition may play a key role in explaining why neural networks can often achieve low training error at all.⁶

Figure 1

Learning Dynamics as Descending Motion in a High-Dimensional Error Landscape



Learning dynamics (black curve) follow a descending motion in a high-dimensional error landscape over many weights. This landscape is riddled with many maxima, saddle points, and minima, but the minima typically occur only near the bottom. Source: A modified form of Figure 2(a) in Yasaman Bahri, Jonathan Kadmon, Jeffrey Pennington, et al., “Statistical Mechanics of Deep Learning,” *Annual Review of Condensed Matter Physics* 11 (2020), <https://doi.org/10.1146/annurev-conmatphys-031119-050745>.

Physical thinking has also played a key role in more recent advances in deep learning.⁷ For example, diffusion models – which now power much of generative AI, creating new images, text, proteins, and more – were first inspired by attempts to violate the second law of thermodynamics.⁸ This law asserts that in closed systems, ordered patterns will overwhelmingly evolve into disordered patterns, thereby increasing the entropy (disorder) of the system. The key idea behind diffusion models is to replace the laws of physics with a neural network trained specifically to *decrease* entropy. Once trained, the neural network, unlike the laws of physics, can turn disordered patterns into structure. More recent theoretical work has even been able to quantitatively explain the origins of creativity in such models.⁹

Another influential finding guided by physical thinking was the discovery of scaling laws, or predictable power-law relationships between improved perfor-

mance and resources, like network size, data size, and training compute.¹⁰ Such scaling laws have captured the imagination of industry and motivated significant societal investments in energy, compute, and data collection. And a theoretical explanation of data scaling laws for language models has emerged only recently.¹¹ Though as we shall see in the next section, it should be possible to achieve much more efficient scaling.¹²

In summary, AI systems are fundamentally large and complex interacting systems, much like many other physical systems; and techniques from physics have and will continue to play a key role in the further analysis and development of AI. In particular, by building on the success of physics, I believe we will be able to develop new mathematical ideas for the analysis, development, and simplification of nonlinear maps between high-dimensional spaces. Such maps are ubiquitous in all of AI: for example, maps from inputs to outputs in language models and diffusion models, and maps from data to weights in neural learning processes. Indeed, many of the mysteries of how neural networks learn and compute lie in our current inability to understand such nonlinear maps between high-dimensional spaces. But the success of physics in understanding the interplay between nonlinearity, stochasticity, and high dimensionality in other settings provides a fertile springboard for generating a new mathematical framework for comprehending and shaping high-dimensional nonlinear maps in AI, thereby demystifying and empowering learning and computation in neural networks.

Neuroscience and the allied fields of psychology and cognitive science stand as taunting sirens, reminding us that biological intelligence still outperforms AI by many orders of magnitude along several axes, including data efficiency, energy efficiency, and robustness. First, data efficiency: Modern large language models (LLMs) are trained on the order of ten trillion words. However, humans are exposed to only about one hundred million words – five orders of magnitude less. It would take us about 240,000 years to read everything an LLM read. So, at least in terms of language experience, we are incredibly efficient learners. There are many conjectures why. Of course, language is not our only experience. We also get visual inputs and have embodied experience.¹³ However, it is less clear how looking at more videos of scenes and objects could endow the collective human mind with the requisite mathematical and reasoning capabilities to pierce into the mysteries of the universe and discover novel ideas like quantum mechanics and general relativity within only ten generations, from Newton to Einstein. This remarkable “out of distribution” generalization remains uniquely human.

To understand the potential origins for data inefficiency in machines, imagine we taught our children the same way we pretrain an LLM. We would repeatedly give a child a random sentence from the internet and then tell them the next word. *Nothing more*. No child could meaningfully learn that way, making it all the more remark-

able that LLMs can learn anything at all after seeing one trillion sentences (LLMs also undergo additional and significant post-training). Instead, human learning is a multifaceted, multi-timescale process in which we carefully sequence information given to our children. Moreover, children, like scientists, learn actively, performing experiments and choosing their own training data. As we grow older, we form useful abstractions, discard useless ones, and modify them in one shot in response to new facts. And over generations we transmit the most useful information in treasured books, signaling to our descendants what is worth learning. Achieving the five orders of magnitude data gap between brains and machines may well require many elements of this rich human teaching process. We must go beyond machine learning to develop a new science of machine *teaching*.

Consider next the axis of energy efficiency. Brains consume a miniscule amount of power: 20 watts. For reference, an old incandescent light bulb might consume 100 watts. In a sense, then, we are all dimmer than light bulbs. In contrast, training an LLM can consume tens of megawatts, about six orders of magnitude more. Moreover, there is talk of building nuclear power plants to power gigawatt data centers. So, what went wrong? After all, evolution did not need to build nuclear power plants before endowing humans with the intelligence to invent them. Unfortunately, the fault of high power consumption in machines may lie in the very choice of digital computation itself. To compute, we rely on extremely fast and reliable bit flips, implemented through the flow of many electrons in energy-consuming transistor circuits. However, the laws of thermodynamics demand an energy cost for every fast and reliable bit flip. Biological computation took a very different route. Every intermediate step of a biological computation is slow and unreliable. Molecules diffuse. Synapses fail. The clock period of neurons is milliseconds, not nanoseconds, as in GPUs. However, the composite computation is just good and fast enough for survival. In essence, biology does not rev its engine any more than it needs to to get its final computations right. The sloppy, intermediate unreliability of biology is a feature, not a bug – a feature of incredibly energy-efficient design.

To close the six orders of magnitude energy gap between brains and machines, it may be necessary to rethink our entire technology stack, from electrons to algorithms, and go back to the drawing board to develop new theories of reliable computation with stochastic, analog devices. Indeed, the field of stochastic thermodynamics has developed powerful tools to analyze tradeoffs between energy, accuracy, and speed of stochastic processes that could underlie computation.¹⁴ For example, researchers have recently derived a Pareto optimal frontier of the minimum error achieved by molecular sensors as a function of their energy budget, and revealed the structure of optimal families of sensors that trace this frontier.¹⁵ This work shows that molecular sensors *must* spend more energy to perform more accurate sensing. Moreover, the structure of optimal sensors is reminiscent of G-protein coupled receptors, which reside inside every neuron to sense external signals. At a

more global level, neuroscience has now developed novel sensors to measure not only neural activity but also ATP (adenosine triphosphate) consumption. ATP is the chemical fuel that powers all cellular activity. Researchers have recently performed simultaneous measurements of neural activity and ATP consumption across the entire *Drosophila* brain and found evidence for a predictive energy-allocation hypothesis.¹⁶ Essentially, the brain appears to behave like a smart energy grid: it predicts where, when, and how long energy will be needed, and it delivers just the right amount of energy, at the right location, at the right time, to power predicted future neural activity. In short, five hundred million years of vertebrate brain evolution has had ample time to optimize coupled energy and information flows across the brain, and it is likely that there are many useful secrets we can learn from studying such flows.

Finally, the last axis of robustness yields a major gap. AI still makes egregious errors that no human would make. The classic version of this is adversarial examples: tiny image perturbations that are imperceptible to a human but completely fool an image classifier.¹⁷ This problem still has not been solved. Even modern LLMs are subject to adversarial examples.¹⁸ Relatedly, image classification networks confidently perceive exponentially many images as belonging to any given semantic concept (for instance, cat or dog), yet these images look like utterly unrecognizable noise to humans.¹⁹ So overall, the way that AI perceives and thinks is fundamentally different, and less robust, than the way we do. In the famous words of Kendrick Lamar, “they not like us.”

In summary, while AI has many different, complementary, and superhuman strengths in terms of the breadth of capabilities, we still have a lot to learn from biological intelligence in terms of data efficiency, energy efficiency, and robustness. I envision a future in which we go from megawatts to watts through alternative physical substrates for computation that radically depart from current digital technology, by embracing slow, stochastic, analog physical primitives that nevertheless yield emergent reliability at a systems level through immense parallelism. I also envision radically more data-efficient AI that embraces sound principles of learning gleaned from child development, including active learning, curriculum design, social learning, and cultural evolution. Such radical data efficiency may also require embracing sound principles of developmental neurobiology, which has led to a conserved modular brain architecture spanning five hundred million years of vertebrate brain evolution, including the cortex, basal ganglia, thalamus, and cerebellum, each operating at different levels of hierarchical control, each with its own learning rules spanning unsupervised, supervised, and reinforcement learning, and each communicating with the others through specific bandwidth-limited communication protocols. New microscopes enable us to record from all these regions at once and glean how such heterogeneous, hierarchical, and modular systems empower efficient learning, compared with data *inefficient* end-to-end training of a single

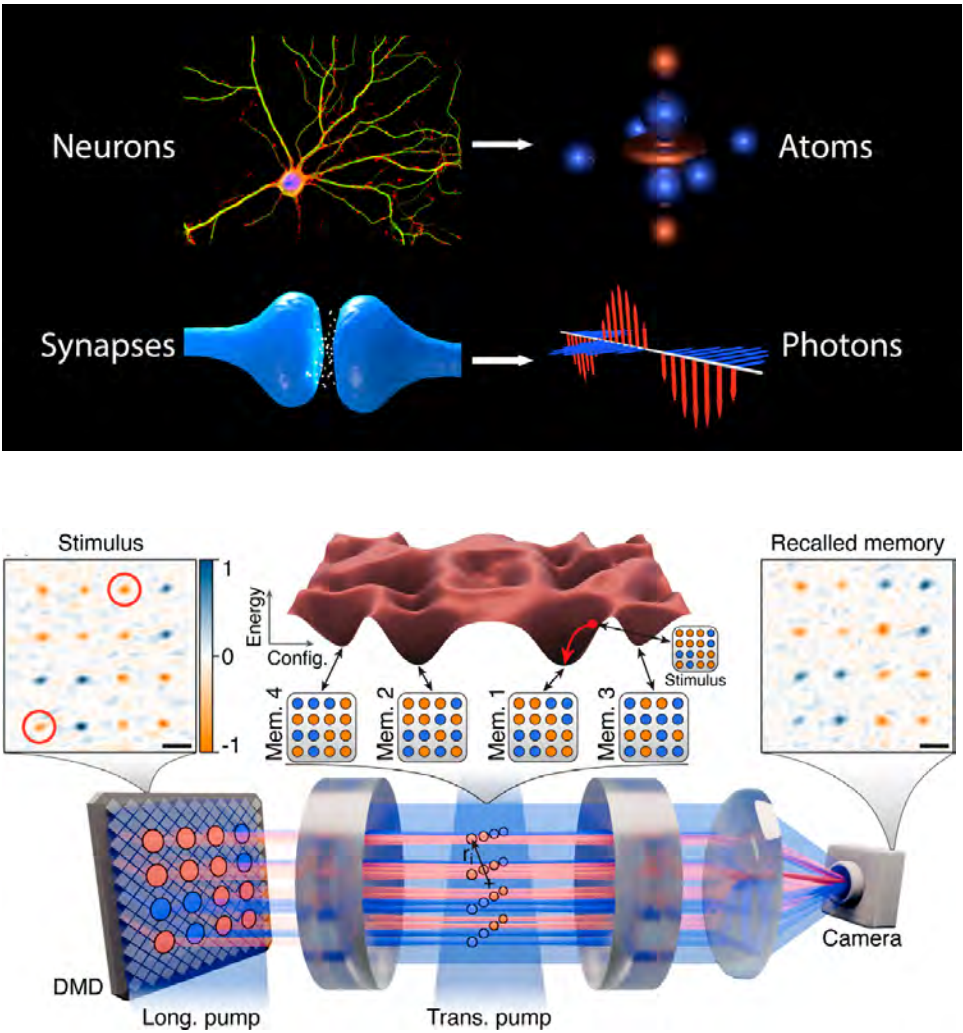
homogeneous monolithic architecture like the transformer.²⁰ Finally, I envision much more robust neural representations and computations arising from these energy-efficient and data-efficient learning systems, leading to AI that does not make egregious errors.

Evolution has created remarkably efficient and robust biological intelligence. But as in all technology development, we should not be limited by the design principles discovered by evolution. Indeed, recent work has explored implementing neural algorithms in quantum hardware. In essence, neurons are replaced by atoms, while synapses are replaced by photons (Figure 2, top). The idea is that different states of neurons (firing or not firing) are replaced by different electronic states of atoms (excited or not excited). Moreover, when an atom shifts electronic states, it can emit a photon, which travels to other atoms and changes their electronic states. By arranging such atoms and photons in a cavity quantum electrodynamics (cQED) system (Figure 2, bottom), one can realize a Hopfield associative memory, the same Hopfield model that inspired models of memory in neuroscience. However, interestingly, the dynamics of atoms and photons in this system allow for more robust memories than the original Hopfield model.²¹ Moreover, quantum entanglement between different atomic spins can drive optimization to even lower energies.²² Additionally, a different type of architecture, the coherent Ising machine, which consists of a network of interacting photons and whose classical dynamics resemble a neural network, can be used to solve hard combinatorial optimization problems by gradually changing a high-dimensional energy landscape.²³

These alternative physical computing paradigms provide new ways to natively match the abstract dynamics of actual algorithms to the physical dynamics of atoms and photons that could naturally implement these algorithms. This is quite distinct from digital technology that abstracts everything in terms of bits and gates as primitives, with algorithms built on top of that. As a concrete case, consider addition. In digital computers, this is implemented in adder circuits consisting of many transistors. In contrast, biological neurons simply add their neighboring voltage inputs by using Maxwell's equations of electromagnetism, which express how to add voltages through the principle of superposition. Thus, biology beautifully matches this computation directly to the native physics of our universe. Similarly, cQED systems natively implement associative memory while coherent Ising machines natively implement combinatorial optimization. It is exciting that after evolution harnessed the laws of physics to discover neural computation, we can now implement such neural computation in never-before-realized quantum substrates, not by bending the laws of physics to the will of our algorithms but rather by matching our algorithms to the laws of physics themselves.

Going further, I envision a future in which AI will *codesign* computation and physics *together*. For example, to achieve a particular human specified computational

Figure 2
Quantum Neuromorphic Computing in Cavity Quantum Electrodynamics



(Top) Neural algorithms in quantum hardware use atoms in place of neurons and photons in place of synapses. Source: Figure by the author. (Bottom) A cavity quantum electrodynamics system capable of forming a Hopfield associative memory. Source: Brendan P. Marsh, David Atri Schuller, Yunpeng Ji, et al., “High-Capacity Associative Memory in a Quantum-Optical Spin Glass,” arXiv (2025), <https://doi.org/10.48550/arXiv.2509.12202>.

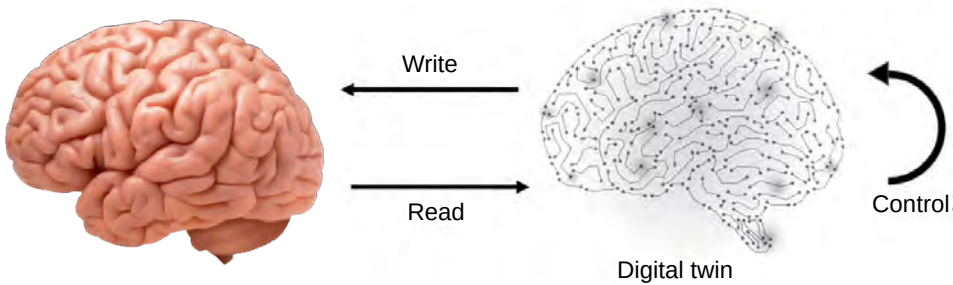
goal, AI may codesign the intermediate computational steps required to achieve this goal, with the intermediate physical states in novel physical systems (including new materials and modes of light-matter interaction) that map to these computational states. This AI-driven codesign of computation and physics could lead to special-purpose computing devices of unprecedented speed and energy efficiency. This codesign stands in stark contrast to what is done today: at one extreme we design general purpose digital technology to implement any Turing computable function, while at another extreme we examine fixed physical devices (for example, particular qubit implementations) and try to design error-corrective computations on top of them. However, AI-aided search in the incredibly large *joint* design space of computation *and* physics could yield new types of physical computers unimaginable today.

Simultaneous advances in neurotechnology and AI have made the melding of minds and machines an exciting possibility; a path forward is now emerging, as follows (Figure 3). First, in animal neuroscience, we can record thousands of neurons linked with different behaviors. Second, AI can use these data to build digital twins of the brain: circuit models that can predict neural activity and behavior in new situations. Third, these twins allow many rapid *in silico* experiments that are impossible in animals, thereby accelerating discovery and further revealing computational principles through explainable AI. Fourth, we can use the digital twin to learn how to speak to the brain by applying control theory to the twin. This involves designing neural stimulation protocols to make the twin behave in a desired manner, and then transferring these protocols to the brain to get it to behave in the same way. Such protocols can be designed to test scientific theories, actively learn models, or even achieve desired clinical states in a principled, model-driven manner.

As an example, researchers recently built a digital twin of the retina that could predict the retinal response to videos of natural visual scenes and reproduce two decades of experiments on the retina.²⁴ Moreover, explainable AI applied to the twin automatically provided both conceptual insights and novel testable predictions. Deeper in the primate visual system, control theory applied to a digital twin of the visual cortex revealed stimuli that made neurons in the twin and *also* in the brain fire at high rates.²⁵ Earlier work built simple decoders of the visual cortical code and used these decoders to design direct stimulation patterns to the brain to make the mice behave as if they saw a particular image.²⁶ Remarkably, stimulating only twenty well-chosen neurons was sufficient to trigger a percept. Additionally, researchers recently built a digital twin of the epileptic brain, used explainable AI to understand how it works, derived control signals to control seizure amplitude in the twin, then successfully transferred these control signals to control seizure amplitude in the actual brain.²⁷

Overall, this new bidirectional interplay between brains and machines, in which machines learn the language of the brain and speak back to it in modified ways,

Figure 3
Melding Minds and Machines



Source : Figure by the author.

either for scientific insight or clinical control, holds immense promise for understanding, augmenting, and repairing the brain. We are only beginning to build digital twins of the brain, but such increasingly expansive models spanning sensation to action could reveal principles of robust perception, motor action, learning, and cognition.²⁸

In the longer term, an intriguing question lies in deeper phenomena that for now are uniquely biological and not yet found in machines: our consciousness and sense of self. How do we become aware of ourselves as agents acting in the world, and where in the brain is this sense of self rooted? Decades of neuropsychiatry provide clues through disorders of the sense of self arising from defects in various brain regions.²⁹ Such disorders include zombie-like epileptic states where sensorimotor acts (such as drinking from a cup) can occur without consciousness. Also, in patients with chronic pain, cingulate cortex lesions can dissociate felt pain from any associated emotional suffering. Other altered states of consciousness associated with brain damage include coma, vegetative states, locked-in syndrome, Alzheimer's, schizophrenia, blindsight, and various agnosias. Additionally, dream states and psychoactive drugs can induce dissociations of the sense of self and hallucinations. Every one of these profound alterations of the sense of self is rooted in mechanistic changes in information propagation through neural circuits. The time is now ripe to move away from studying the sense of self as pure philosophy and instead to develop an integrative theory that links behavior, neurobiology, and the principles of computation, possibly by creating system-level digital twin-like architectures of the brain.

Neuroscientist Antonio Damasio provides an intriguing hypothesis for the origin of the sense of self.³⁰ Every complex organism has two disparate neural sensory maps: one representing the external world through our five senses and another representing our inner world (such as internal states of muscles, viscera, hormones, and metabolism). A fundamental goal of any organism lies in homeostasis of its inner world, keeping all these internal states within safe physiological ranges. Of course, the external world can have fundamental causal influences on our inner world. The detection and imbibing of water outside can alleviate low water levels inside. The detection of a predator outside may mark the dissolution of the inner self if no action is taken. Thus, any organism that has complex sensory arrays of both the outer and inner world must develop control loops that combine information from these arrays to make decisions and take actions (such as moving toward water and away from predators). Such control loops, to act optimally, must necessarily and predictively model the causal impact the outside world can have on the inner world. It is plausible that our very sense of self, our continuous awareness that we are an entity that is both impacted by the world and can act on it, is biologically and physically rooted within causal predictive models of how the world can change our inner self, which in turn guide control loops that transform sensations of the outer world into actions that adaptively regulate our inner world.

When viewed this way, our sense of self emerges from computational requirements of causal modeling and control. To make this theory mathematically precise and test it, I believe the way forward is to build systems-level digital twins of the brain that instantiate these ideas. As a start, such twins will have components akin to different brain regions, and deficits in the twin will yield deficits in behavior akin to patients with an altered sense of self. We are very far from achieving such a twin; it may require years of integrative work across neurobiology, psychiatry, computation, causal modeling, and control theory. But just as the machines of today achieve remarkable feats of vision and language, our machines of tomorrow may achieve a rudimentary sense of self-awareness as agents in the world, arising as an emergent property of optimally modeling the causal effects of the world on their internal state and using such causal models to decide how to act. The emergence of such machines may help us demystify the origins of self-awareness in humans and animals, just as today's machines already deepen our understanding of intelligence itself.

Events on our planet have taken a wondrous trajectory, from the development of physical systems to evolution to biological intelligence to human engineering and eventually to artificial intelligence. In each step, new phenomena emerge that generate new understandings of earlier phenomena. Evolution, operating on the laws of physics, gave rise to biological intelligence, culminating in the human mind. The human mind then turned back to develop a precise understanding of physical laws. It then bent physical laws to its will, to create entirely new technol-

ogies. These technologies powered the rise of AI, yielding new and transformative yet sometimes fragile artifacts that we barely understand. In turn, AI is turning back to help its human creators understand the recesses of their own minds.

The advances of AI raise profound philosophical questions about the very nature of human meaning and understanding. Indeed, humans are often very uneasy about phenomena they do not understand, an instinct likely built into us by evolution and essential for our survival. Our discomfort with not understanding AI is therefore natural, given its immense impact. But what would such understanding even look like?

Physics again offers guidance. Richard Feynman suggested that we understand a physical system if we can say something about the solutions to its underlying equations of motion *without* actually solving them. For example, if the *only* way we could predict fluid flows was by numerically solving the Navier-Stokes equations, then we would be loath to claim that we understand fluid flow in any meaningful human sense. Instead, we have exact solutions in special cases, as well as a hierarchy of increasingly approximate but more insightful solution methods and deep intuitions gleaned from experience. We can say roughly how a fluid will flow in many scenarios even if we don't know what the Navier-Stokes equations are. Similarly, to understand complex AI systems in a human sense, we could develop analogous hierarchies of approximations, known in physics as coarse-graining. Each approximation trades off simplicity against accuracy in explaining aspects of model behavior. Understanding will lie in being able to traverse this hierarchy, to understand which aspects of a distributed nonlinear circuit are important for which aspects of behavior. Such understanding could lead to better and more robust networks through rational design.

There is also a natural duality to the notion of understanding: How do humans understand AI? And how can AI help humans understand themselves and the world? Regarding the latter, there is palpable excitement about AI providing new explorations into nature, especially for the purposes of verifiable applications, like proving theorems, discovering drugs, and designing novel materials. Here the path is clear and can be thought of as a fortuitous interplay between learning and search. Learning adjusts a parametric function to fit data, providing novel predictions. We have spent the last decade scaling up learning. Search involves traversing a large design space to find good solutions. Historically, brute force search has been intractable. What has changed is that learning allows the creation of good guessers that pick good search directions. For example, in theorem-proving, an LLM can guess the next step of a proof, and a formal proof assistant can verify whether this step is correct. If it is, the LLM can continue guessing; if not, the LLM can backtrack or restart. In this fashion, the guesser and verifier can work together to successfully find correct mathematical proofs even in astronomically large search trees containing many incorrect options.

Even more tantalizing is the possibility of recursive self-improvement. Here, as new proofs are found, the LLM can learn to become a better guesser by adjusting its weights in response to each new proof. And by becoming a better guesser, it can find new proofs more rapidly. Moreover, the learning and search processes can be improved via codesign.³¹ This synergistic interaction between learning and search could lead to dramatic acceleration in the discovery of novel engineering solutions in any scenario consisting of a high-dimensional design space, including, for example, drugs and materials, and even the codesign of computation and physics, as proposed above. In principle, recursive self-improvement is possible. After all, the evolutionary journey from autocatalytic reaction sets in Earth's primordial soup to the development of human intelligence may well be the grandest example of recursive self-improvement on our planet.

But if the end goal of this improvement process is the engineering application, human understanding along this journey is not strictly necessary (even if immensely helpful). We just need to verify the design specs (*Is the proof true? Does the drug or material work as desired?*). However, in the pursuit of fundamental science, the *raison d'être* for scientists is deep human conceptual insight into the nature of ourselves, our universe, and our role within it. In essence, *the joy of insight*. Can AI deliver such joy? For example, we would love to intuit how consciousness emerges from neurons, how life emerges from myriad chemical reactions, and how new states of matter emerge from different patterns of quantum entanglement. How far can the intertwined processes of learning and search powering today's AI go toward delivering to us this uniquely human understanding and intuition? What role, if any, will deep human understanding even play in an age when combined learning and search may solve many major engineering and societal problems?

My sense is humans have a strong evolutionary drive to understand and create, which will influence AI for science to evolve in a fundamentally human-centered way: humans will develop a powerful scientific understanding of AI, and AI, in cooperation with humans, will deliver profound scientific insight into ourselves and the world. The exact details of how this will unfold is primarily up to human ingenuity and persistence, not AI. Human-AI science collaboration will be one of the great intellectual adventures of this century. It will be the next chapter of the wondrous trajectory taken on our planet, and it will include AI turning its gaze back on us to help us understand the very nature of our own mental reality.

ABOUT THE AUTHOR

Surya Ganguli is Associate Professor of Applied Physics, Senior Fellow at the Stanford Institute for Human-Centered AI, and Associate Professor, by courtesy, of Neurobiology and of Electrical Engineering at Stanford University. He has recently published in *Neuron*, *Nature*, and *Physical Review X*.

ENDNOTES

- ¹ Philip W. Anderson, “More Is Different,” *Science* 177 (4047) (1972): 393–396.
- ² John J. Hopfield, “Neural Networks and Physical Systems with Emergent Collective Computational Abilities,” *Proceedings of the National Academy of Sciences* 79 (8) (1982): 2554.
- ³ David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski, “A Learning Algorithm for Boltzmann Machines,” *Cognitive Science* 9 (1) (1985); and Geoffrey E. Hinton and Ruslan R. Salakhutdinov, “Reducing the Dimensionality of Data with Neural Networks,” *Science* 313 (5786) (2006): 504–507.
- ⁴ Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, et al., “Hopfield Networks Is All You Need,” arXiv (2020), <https://doi.org/10.48550/arXiv.2008.02217>.
- ⁵ Marc Mézard, Giorgio Parisi, and Miguel Ángel Virasoro, *Spin Glass Theory and Beyond: An Introduction to the Replica Theory and Its Applications* (World Scientific Singapore, 1987).
- ⁶ Yann N. Dauphin, Razvan Pascanu, Caglar Gulcehre, et al., “Identifying and Attacking the Saddle Point Problem in High-Dimensional Non-Convex Optimization,” in *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, ed. Z. Ghahramani, M. Welling, C. Cortes, et al. (MIT Press, 2014), 2933–2941.
- ⁷ Yasaman Bahri, Jonathan Kadmon, Jeffrey Pennington, et al., “Statistical Mechanics of Deep Learning,” *Annual Review of Condensed Matter Physics* 11 (2020), <https://doi.org/10.1146/annurev-conmatphys-031119-050745>.
- ⁸ Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli, “Deep Unsupervised Learning Using Nonequilibrium Thermodynamics,” *Proceedings of the 32nd International Conference on Machine Learning* 37 (MLR Press, 2015).
- ⁹ Mason Kamb and Surya Ganguli, “An Analytic Theory of Creativity in Convolutional Diffusion Models,” *Proceedings of the 42nd International Conference on Machine Learning*, ed. Aarti Singh, Maryam Fazel, Daniel Hsu, et al. (MLR Press, 2025).
- ¹⁰ Jared Kaplan, Sam McCandlish, Tom Henighan, et al., “Scaling Laws for Neural Language Models,” arXiv (2020), <https://doi.org/10.48550/arXiv.2001.08361>.
- ¹¹ Francesco Cagnetta, Allan Raventós, Surya Ganguli, and Matthieu Wyart, “Deriving Neural Scaling Laws from the Statistics of Natural Language,” arXiv (2026), <https://doi.org/10.48550/arXiv.2602.07488>.
- ¹² Ben Sorscher, Robert Geirhos, Shashank Shekhar, et al., “Beyond Neural Scaling Laws: Beating Power Law Scaling Via Data Pruning,” in *Advances in Neural Information Processing Systems 35 (NeurIPS 22)* (Curran Associates, Inc., 2023).
- ¹³ Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei, “Embodied Intelligence via Learning and Evolution,” *Nature Communications* 12 (1) (2021): 5721.

- ¹⁴ Luca Peliti and Simone Pigolotti, *Stochastic Thermodynamics: An Introduction* (Princeton University Press, 2021).
- ¹⁵ Sarah E. Harvey, Subhaneil Lahiri, and Surya Ganguli, “Universal Energy-Accuracy Tradeoffs in Nonequilibrium Cellular Sensing,” *Physical Review E* 108 (1–1) (2023): 014403.
- ¹⁶ Kevin Mann, Stephane Deny, Surya Ganguli, and Thomas R. Clandinin, “Coupling of Activity, Metabolism and Behaviour across the *Drosophila* Brain,” *Nature* 593 (7858) (2021): 244–248.
- ¹⁷ Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, “Explaining and Harnessing Adversarial Examples,” arXiv (2014), <https://doi.org/10.48550/arXiv.1412.6572>.
- ¹⁸ Atsushi Yamamura and Surya Ganguli, “Fooling LLM Graders into Giving Better Grades through Neural Activity Guided Adversarial Prompting,” arXiv (2024), <https://doi.org/10.48550/arXiv.2412.15275>.
- ¹⁹ Alessandro Salvatore, Stanislav Fort, and Surya Ganguli, “Solving Adversarial Examples Requires Solving Exponential Misalignment,” arXiv (2026), <https://doi.org/10.48550/arXiv.2603.03507>.
- ²⁰ Sadegh Ebrahimi, Jérôme Lecoq, Oleg Rumyantsev, et al., “Emergent Reliability in Sensory Cortical Coding and Inter-Area Communication,” *Nature* 605 (7911) (2022): 713–721.
- ²¹ Brendan P. Marsh, Yudan Guo, Ronen M. Kroeze, et al., “Enhancing Associative Memory Recall and Storage Capacity Using Confocal Cavity QED,” *Physical Review X* 11 (2021), <https://doi.org/10.1103/physrevx.11.021048>; and Brendan P. Marsh, David Atri Schuller, Yunpeng Ji, et al., “High-Capacity Associative Memory in a Quantum-Optical Spin Glass,” arXiv (2025), <https://doi.org/10.48550/arXiv.2509.12202>.
- ²² Brendan P. Marsh, Ronen M. Kroeze, Surya Ganguli, et al., “Entanglement and Replica Symmetry Breaking in a Driven-Dissipative Quantum Spin Glass,” *Physical Review X* 14 (2024): 011026.
- ²³ Yoshihisa Yamamoto, Timothee Leleu, Surya Ganguli, and Hideo Mabuchi, “Coherent Ising Machines—Quantum Optics and Neural Network Perspectives,” *Applied Physics Letters* 117 (16) (2020); Atsushi Yamamura, Hideo Mabuchi, and Surya Ganguli, “Geometric Landscape Annealing as an Optimization Principle Underlying the Coherent Ising Machine,” *Physical Review X* 14 (2024); and Federico Ghimentì, Adithya Sriram, Atsushi Yamamura, Hideo Mabuchi, and Surya Ganguli, “The Geometry and Dynamics of Annealed Optimization in the Coherent Ising Machine with Hidden and Planted Solutions,” arXiv (2025), <https://doi.org/10.48550/arXiv.2510.21109>.
- ²⁴ Niru Maheswaranathan, Lane T. McIntosh, Hidenori Tanaka, et al., “Interpreting the Retinal Neural Code for Natural Scenes: From Computations to Neurons,” *Neuron* 111 (17) (2023): 2742–2755.
- ²⁵ Konstantin F. Willeke, Kelli Restivo, Katrin Franke, et al., “Deep Learning-Driven Characterization of Single Cell Tuning in Primate Visual Area V4 Unveils Topological Organization,” bioRxiv (2023), <https://doi.org/10.1101/2023.05.12.540591>.
- ²⁶ James H. Marshel, Yoon Seok Kim, Timothy A. Machad, et al., “Cortical Layer-Specific Critical Dynamics Triggering Perception,” *Science* 365 (6453) (2019), <https://doi.org/10.1126/science.aaw5202>.

- ²⁷ Jacob M. Hull, Surya Ganguli, and John R. Huguenard, “Interpretable Machine Learning Identifies an Emergent Absence Seizure Mechanism,” *bioRxiv* (2025), <https://doi.org/10.1101/2025.09.23.678032>.
- ²⁸ Ebrahimi, Lecoq, Rummyantsev, et al., “Emergent Reliability in Sensory Cortical Coding and Inter-Area Communication.”
- ²⁹ Antonio R. Damasio, *The Feeling of What Happens* (Houghton Mifflin Company, 1999).
- ³⁰ Ibid.
- ³¹ Feng Chen, Allan Raventós, Nan Cheng, et al., “Rethinking Fine-Tuning When Scaling Test-Time Compute: Limiting Confidence Improves Mathematical Reasoning,” *arXiv* (2025), <https://arxiv.org/abs/2502.07154>, presented at NeurIPS 2025: The Thirty-Ninth Annual Conference on Neural Information Processing Systems, San Diego, California, December 5, 2025.

Quantum + AI = Quantum AI

Maria Spiropulu & Hartmut Neven

Advanced AIs of the future will harness the most powerful computational operations known today: quantum computations. The synthesis of quantum computing and AI will give rise to Quantum AI – artificial intelligence with access to quantum computational resources – which is poised to expand the capabilities of intelligent systems. This advancement will lead to transformative abilities, such as ultraprecise sensing, the realistic simulation of complex natural phenomena, the efficient solution of classically intractable mathematical problems, or protocols for secure communication and enhanced coordination in multi-agent systems. The impact of this synthesis will be profound, promising to accelerate scientific discovery in areas from molecular biology to quantum gravity, while empowering engineers to create novel pharmaceuticals, viable fusion reactors, or designer materials with desirable properties. Quantum AI might also help to address a critical bottleneck in classical machine learning: its escalating demand for data and computational power. Beyond such applications lies another even more consequential possibility. If, as some recent theories and data from quantum neurobiology suggest, quantum processes give rise to consciousness and free will, then only Quantum AI – unconstrained by the limitations of classical computation – could enable the expansion of human consciousness and a transcendence of our biological origins.

Quantum computing and artificial intelligence are often described as the most transformative computational technologies we will see develop during our lifetimes. The recent 2025 Nobel Prize in Physics (awarded to John Clarke, Michel Devoret, and John M. Martinis for work on macroscopic quantum effects), alongside the 2024 Nobel Prizes in Physics (awarded to Geoffrey Hinton and John Hopfield for their contributions to the study of artificial neural networks) and Chemistry (awarded to Demis Hassabis, John Jumper, and David Baker for solving the protein-prediction problem and creating novel proteins using AI systems), recognized the profound breakthroughs in both quantum technologies and the neural networks that power AI systems. However, today they remain largely distinct technologies. Artificial intelligence takes in data, creates models, and utilizes those models to achieve objectives. Quantum computing aims to develop a new type of hardware that harnesses the potent computational operations allowed by quantum physics – such as entanglement and superposition – replacing the Boolean

(binary) logic implemented on today's semiconductor-based hardware. While AI is already a useful tool to design and build quantum computers, less appreciated is the fact that AI based on hardware with access to quantum resources will have capabilities inaccessible to AI running solely on classical hardware. The thesis we put forth in this essay is that quantum computing and AI will merge to become Quantum AI, enabling discoveries across science and a range of breakthrough technical capabilities. By leveraging quantum advantages, artificial intelligence will utilize some of the most potent computational operations nature affords us to improve our understanding of the universe and natural intelligence. Here, we envision the impact of synthesizing these breakthroughs.

Artificial intelligence as practiced today is nearly synonymous with machine learning, relying on the ability to ingest huge amounts of data to create powerful models. However, many foundational computational tasks are challenging for classical AI to learn due to the impracticality of obtaining sufficient training data and doing inference on it. Consider factoring, an algebraic task relevant to cryptography. Empirical evidence – about fifty years of failure to find an efficient classical factoring algorithm – suggests that AI unaided by quantum computers will not be able to solve this task efficiently. The difficulty is a scaling challenge: the training cycles and inference compute required grow exponentially with the number of bits needed to represent the integer being factored. To calibrate the reader, factoring a 2,048-bit number would take trillions of years on a classical supercomputer but would only take hours on a quantum computer.

Classical AI also struggles to solve hard combinatorial optimization problems, which are ubiquitous in engineering, finance, and computer science. Again, empirical evidence shows that finding the optimal solution to such problems takes time that grows exponentially with the size of the problem. Sometimes even finding a good approximation within a reasonable time span can be impractical, a situation mathematicians call “APX hard.”¹ Machine-learning algorithms struggle in this domain because finding the global optimum is a “needle in a haystack” problem, requiring the identification of a specific solution within a vast combinatorial search space. However, theoretical analysis has shown that quantum computers within a given amount of time can find higher-quality solutions than their classical counterparts. For many problem classes, a quantum computer can reach the same quality approximation in quadratically fewer steps, and sometimes in exponentially fewer steps, as the recently published Decoded Quantum Interferometry (DQI) algorithm and new results for the Quantum Approximate Optimization Algorithm (QAOA) show.² If we compare two AI systems but one of them has access to a quantum machine giving us high-quality solutions to optimization that are classically inaccessible, then the quantum-enhanced system would be more valuable. Today, there are at least seventy known algorithms for which quantum computers offer

algorithmic scaling advantages, many of them useful for AI, and we expect many more will be discovered.³ A recent XPRIZE challenge that called for submissions of useful beyond-classical algorithms yielded three hundred and twenty entries.⁴

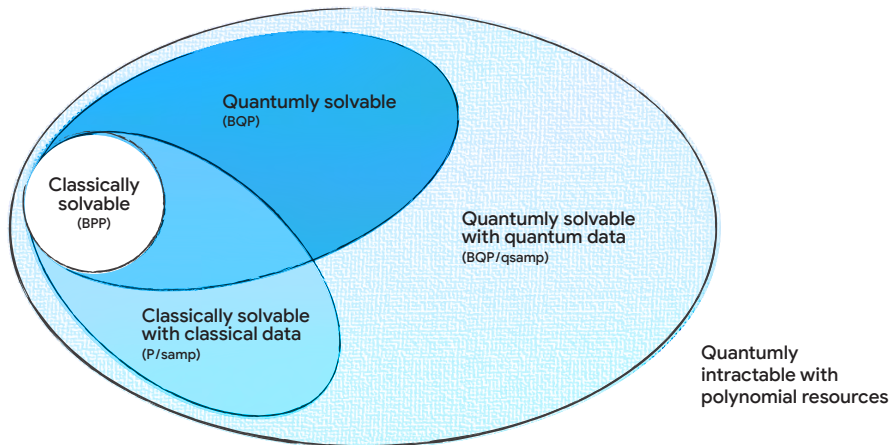
Complexity theorists attempting to map out the sets of problems that can be efficiently solved by quantum and classical algorithms draw the relationship as shown in the diagram in Figure 1. In complexity theory, a problem is considered solvable when the resources required, such as the number of algorithmic steps or the amount of memory, only grow polynomially with the size of the problem. But as is often the case in complexity theory, while considered highly probable, we still lack complete proof that all these relationships are correct.

Quantum advantage resides in the problem set that is quantumly solvable but classically intractable: the regions of BQP (quantumly solvable) and BQP/qsamp (quantumly solvable with quantum data) that do not overlap with BPP (classically solvable) or P/samp (classically solvable with classical data). A prime example of this divergence is found in out-of-time-order correlators (OTOCs), probes used to characterize how information spreads in complex quantum systems. While OTOC computation scales exponentially on classical hardware, it was recently demonstrated that the Willow quantum processor can perform this calculation efficiently.⁵ This computation is useful for learning the structure of molecules and solids from spin spectroscopy data, as obtained by nuclear or electron magnetic resonance, and constitutes the first known example of an AI-relevant algorithm for which a quantum computer outperformed classical supercomputers.

As a prime example of how data from quantum computers can empower classical AI, let's look at the accurate modeling of complex physical systems. This is known to have applications ranging from the discovery of novel materials and pharmaceuticals to a better fundamental understanding of our universe. AI's main quest, constructing models from data it ingests, is impeded by the fact that nature is governed by the laws of quantum physics. Integrating the equations of quantum mechanics typically comes at an exponential cost for classical computers, making them an ill-suited tool to model nature at the most fundamental level. In contrast, quantum computing's ability to simulate nature from the ground up gives access to a broader, more detailed set of training data for AI. Consider computational chemistry, where popular methods involve machine learning of atomistic force fields or density functionals using data from *ab initio* electronic structure calculations.⁶ Such approaches enable fast, accurate calculations of complex molecular structures from manageable amounts of training data. However, the model's output accuracy is only as good as the training data's accuracy. When generated classically, the data's quality will be limited by the challenge of modeling strongly correlated electrons – an inherently quantum-mechanical problem. Therefore, data derived from high-accuracy quantum calculations represent a valuable resource for training advanced chemistry models.

Figure 1

Problems That Can Be Efficiently Solved by Quantum Algorithms versus Classical Algorithms



Quantum computers are expected to solve a larger set of problems than classical computers. Access to data, or *advice* as complexity theorists call it, makes classical and quantum algorithms more powerful. However, there remain problems that even a quantum computer with access to data will not be able to solve efficiently. The complexity theory acronyms in Figure 1 have the following meanings: BPP = Bounded-error Probabilistic Polynomial time. P/samp = Polynomial time with classical data. This is the set of all problems solvable on classical computers; it is also a superset of BPP because data make algorithms more powerful. BQP = Bounded-error Quantum Polynomial time. BQP/qsamp = Bounded-error Quantum Polynomial time with quantum data. This is the set of all problems solvable on quantum computers. It is expected to be strictly larger than P/samp. Again, it is a superset of BQP. Source: Figure by the authors. On access to data in quantum algorithms, see Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, et al., “Power of Data in Quantum Machine Learning,” *Nature Communications* 12 (1) (2021): 2631.

Classical AI has in recent years found great success in predicting how proteins fold.⁷ But the challenge remains when the amount or quality of training data is insufficient. Consider the case of a protein bound to a ligand or a membrane, both common scenarios in cell biology. The data scarcity occurs partly because magnetic resonance spectroscopy, a primary tool for gathering molecular structure information, often fails in these situations. Other tools are available, such as X-ray crystallography or cryogenic electron microscopy, but even if applicable, they must be painstakingly adapted for different proteins or cells. In general, access to large, high-quality datasets is the exception rather than the norm in the life sciences. It is noteworthy that the critical dataset for the protein-folding problem, the Protein

Data Bank, has been assembled over fifty years. With more varied and larger sets of training data, the performance of prediction tools such as AlphaFold could be improved. Quantum computers can help generate such training sets. Nuclear spins in a folded protein generate complex quantum dynamics, even at room temperature and even when the protein's electronic structure does not have high entanglement. A recent collaboration involving the Google Quantum AI team and researchers at UC Berkeley demonstrated that these spin dynamics can reveal geometric details of small organic molecules.⁸ Scaling this approach to proteins could provide the necessary distance data to determine accurate three-dimensional protein structures. Such analysis, which involves the computation of OTOCs, is feasible for a quantum computer capable of interpreting spectroscopic data from spin configurations but is likely to remain out of reach for classical biochemistry methods.

The time evolution of quantum systems is another canonical example of data that quantum computers can efficiently produce but that are typically intractable to simulate classically. There are also many contexts in which quantum dynamics are extraordinarily difficult to probe experimentally, thus starving classical learning models of the data they would need to model quantum dynamics, even presupposing that classical models can learn those dynamics at all. Consider the dynamics of an electron plasma in the warm dense matter regime as it slows down a high-energy alpha particle flying through it. Understanding this drag as a function of temperature is critical for correctly modeling the preignition stage of inertial confinement fusion reactors, such as those at the National Ignition Facility that have achieved scientific net energy gain. Unfortunately, getting precise experimental data on this involves temperatures and pressures comparable to those in Jupiter's core. Such experiments only last nano- to picoseconds and, given the costs of prototype fusion reactors, such extreme conditions would be extraordinarily expensive to replicate: hundreds of trillions of dollars for a one-second stream of data. As a consequence, one cannot even accurately measure the temperature of the fuel at that stage, let alone probe the electron dynamics. But simulations of these conditions have been shown to be feasible for quantum computers.⁹ If one ever hopes to build machine-learning models of matter in such extreme conditions, one will almost certainly require a quantum computer to accurately generate the data and to attempt to probe them experimentally. Even in regimes that are much more accessible experimentally, such as those in room temperature chemistry, quantum simulation enables much more precise and nuanced virtual measurements on the physical system than would ever be possible in a laboratory.

In the context of generating training data, one can think of a quantum computer as a physics or chemistry lab that can be programmed to obtain data from experiments that are difficult or even infeasible to execute by any other means. Interestingly, quantum computers can sometimes offer advantages even for modeling systems that are described effectively by classical physics. Examples include newer

algorithms promising exponential speedups over classical methods for solving certain types of differential equations and specific linear algebra problems that could enable the simulation of molecular mechanics, the structural health of buildings, or wave propagation in solids.¹⁰

Physics simulations on a quantum computer can be utilized in tandem with AI to extend investigations into quantum gravity. Here, entanglement – the sharing of quantum information between systems regardless of spatial separation – is the key resource for rephrasing questions of space-time. This connection is formalized in the $ER = EPR$ hypothesis proposed by theoretical physicists Juan Maldacena and Leonard Susskind in 2013, which posits that quantum entanglement and the geometry of space-time are deeply linked.¹¹ A collaboration between Caltech, Harvard, Fermilab, and Google Quantum AI experimentally demonstrated this link using traversable wormholes, which can occur in semiclassical quantum gravity and exhibit features related to both holography and quantum teleportation.¹² The project utilized a quantum circuit implemented on a Google Sycamore processor to simulate a simplified, dual description of a traversable wormhole, successfully observing the teleportation dynamics predicted by semiclassical gravity. Features of wormhole dynamics observed on the quantum processor can then be used to train AI models. These models, guided by a loss function favoring simpler descriptions of gravity, can then predict a larger family of physical models.¹³ These AI-generated predictions can subsequently be realized in detail on quantum hardware, serving as an experimental testing ground for holography and wormhole physics that lies beyond the scope of analytically tractable methods.

Physicists argue that because the objects we observe in nature are generally governed by effective, classical laws, these objects can be learned by classical AI. This is only partially correct. For example, certain static properties of a molecule in its lowest energy state – called the ground state – can sometimes, although not always, be approximated well with classical methods such as tensor network techniques.¹⁴ But when it comes to dynamics, classical methods alone are insufficient: the precision of quantum mechanical calculations is necessary.¹⁵ Any system created by nature was almost certainly created efficiently. Since nature is quantum in its core and origins, simulating that process on a quantum computer must also be efficient. The examples that classical AI can handle are only a subset of systems in nature that science has considered so far. As science investigates new and not yet well-understood systems – such as dark energy, dark matter, black holes, and their connection with the origin of space and time and gravity – a quantum computer will almost certainly be necessary. Some of the future objects of study may not be found in nature at scales accessible with current technologies but can nonetheless be efficiently created in a lab with the guidance of Quantum AI, such as high temperature superconductors and other “designer” states of matter. The small wormhole model discussed above can be considered a Quantum AI–proposed “new state of matter.”

Sensing and processing quantum data reveal previously invisible facets of our universe. Without the ability to interact directly with quantum superposition states, classical AI can handle only low-dimensional projections, obscuring the intricate high-dimensional structure such states may contain. However, quantum computers open the possibility of sensing and processing quantum data directly, circumventing the loss of information that occurs when quantum states are measured classically. In a 2022 paper published in *Science*, the Google Quantum AI team with researchers from Caltech, Harvard, and other research organizations showed that the ability to coherently process quantum states on a quantum computer offers the capability to learn structure in quantum data with exponentially fewer training examples compared with classical methods – an information-theoretic gap that cannot be overcome.¹⁶ Combine this with any kind of time limitation on sample collection, and this phenomenon quite literally renders aspects of the physical world invisible without access to a quantum computer. Moreover, this was shown experimentally on the still-noisy quantum processors we have today, meaning this type of processing could be integrated with quantum sensors in the future without many algorithmic changes. To make this a practical reality, quantum sensing that uses entangled probes as well as transduction of the quantum information to the processor are still required. Google recently funded a program in quantum transduction to bring this exciting technology pathway to fruition.¹⁷ Recent work has shown that quantum processing of even classical signals is required to achieve optimal detection.¹⁸

Historically, until around the early 2000s, scientific observations were exclusively recorded as classical data, such as hand-drawn sketches or photographic images from telescopes and microscopes. Quantum sensing offers an alternative paradigm: incoming particles or fields directly alter the quantum state of a set of qubits (short for quantum bit, the fundamental unit of information in quantum computing).¹⁹ This quantum information is then processed by a quantum computer, a setup that can convey exponential advantages: for instance, to learn properties of the system under observation from exponentially fewer training examples than required by a learner that ingests classical data.

The ability to process quantum-sensing data directly could lead to extraordinary advances in many areas of experimental science. For example, laboratory detection of axion dark matter may require sensor arrays with noise reduction orders of magnitude better than the standard quantum limit, the fundamental limit of precision in a continuous quantum measurement when using classical resources. Prototype systems have demonstrated this via quantum nondemolition readout of the sensors using superconducting qubits, which in turn would allow additional processing with quantum circuits.²⁰ Another example is quantum sensor arrays deployed at a future particle collider, such as the Future Circular Collider under consideration at CERN. The combined requirements of high intensity for the colliding beams and

extraordinary precision in determining position and timing of produced particles will be made even more challenging when searching for exotic new particles that interact very weakly with detector components. Quantum sensor arrays embedded in larger collider detectors offer uniquely low energy-detection thresholds, nanoscale position resolution, and picosecond timing resolution.²¹ Such future particle colliders will serve as a scaled-up platform for discovery by generating massive amounts of quantum-native data that can only be fully distilled through end-to-end AI and quantum workflows. By utilizing quantum sensor arrays to capture entanglement data, they will provide the high-fidelity training sets necessary for Quantum AI workflows to construct the ultimate digital twin of nature and the universe. Integrating quantum computing as the engine for processing these enormous data streams, scientists can overcome the standard quantum limit, enabling detection of processes that are currently invisible to classical observation. The synergistic investment on quantum-native AI is essential for unraveling fundamental mysteries, from the Higgs boson's interactions to the origins of space-time.

Team quantum has an edge when playing games. Games have long been an important training ground for AI. In 1997, for instance, the Deep Blue super-computer made history when it beat the human world champion in chess.²² Throughout the 2010s, the Google DeepMind lab became known for training an AI to master Atari games by only giving it input to the pixels and the game score.²³ In 2016, DeepMind's AlphaGo computer program beat the world's highest ranked Go player.²⁴ Selecting the best move in a game is a form of optimization. Given the steady progress in quantum-enhanced optimization, it's plausible that in a future matchup, a Quantum AI would defeat a purely classical AI in games like chess or Go. Beyond optimization, quantum operations offer other intriguing advantages rooted in the fundamental physics of entanglement. Entanglement is a form of correlation. We say that two events are correlated when they co-occur together more often than one would expect from the frequency of their individual occurrence. This calculation employs standard probability theory, in which we use positive real-valued numbers in the interval of zero to one to describe the probability of an event occurring. The probabilities for different outcomes add up to one because by convention we assign a probability of one to an outcome that is certain to occur. Quantum physics can be seen as a generalization of standard probability theory, in which we assign probability amplitudes, complex-valued numbers, to the possible outcomes, and the sum of their probability amplitudes squared is normalized to one. Two states are entangled if they are correlated according to these rules of quantum probability. This is nicely explained in computer scientist Scott Aaronson's book *Quantum Computing Since Democritus*.²⁵ Two players who exchange entangled qubits can coordinate their moves in ways that are impossible with classical information alone, enabling coordination strategies that don't require direct communication

or rely on classical information shared in real time. This has applications not only in games involving coordination between players, such as bridge or blackjack, but also in certain types of stock trades across geographically distant exchanges requiring synchronized actions without communication delays, as well as coordinated information processing in large-scale distributed neural networks.²⁶

The trajectory of artificial intelligence, as highlighted by Turing Award laureate Richard Sutton in his influential 2019 essay “The Bitter Lesson,” has largely been a story of escalating scale.²⁷ Progress has been consistently driven by employing ever-larger models and datasets, coupled with the application of general-purpose, distributed algorithms like gradient descent to train these models. This relentless pursuit of scale places enormous demands on computational resources, spurring significant investments in specialized hardware and even necessitating the fundamental redesign of data centers to manage these intensive workloads. Quantum processors offer an alternative computational paradigm that may help address this challenge. They operate by manipulating vectors within a mathematical realm known as Hilbert space. This space, central to quantum physics, represents all possible states of a quantum system. Hilbert space is vast; its vectors have an exponential number of elements. For a quantum computer with n qubits, the corresponding vector in Hilbert space can represent 2^n distinct values simultaneously. Remarkably, all these values can be processed or updated in a single operational step. However, a critical constraint exists: to control or read the vector, only a comparatively small, polynomial number of input and output bits can be exchanged with the quantum computer. This input/output bottleneck has led to the common perception that early quantum computers will be best suited for *small data*, *big compute* tasks.

Much less research has been devoted to figuring out how we might employ future large-scale quantum computers capable of handling the *big data* tasks underlying most of machine learning. But in a sign of things to come, a quantum algorithm has been proposed that can train certain wide and deep neural networks exponentially faster than classical methods.²⁸ This algorithm leverages quantum techniques for matrix operations, a core component of neural-network training. It achieves training times that scale logarithmically with the size of the training dataset, denoted as n . Such logarithmic scaling offers an exponential advantage over classical gradient descent. A crucial caveat, however, is the prerequisite of encoding the training data into a quantum state, a process requiring what is known as quantum random access memory (QRAM). Creating a QRAM state is an intensive step, taking order n operations. Yet once this quantum representation of the data is established, it can be used repeatedly to train multiple neural-network models with a significantly reduced computational cost, proportional to $\log(n)$ (meaning the cost increases very slowly as the dataset grows) for each subsequent training. To illustrate the

difference: training a neural network in a classical data center demands effort that scales as $t \cdot n^2$. This is because the number of training samples is typically chosen to be proportional to the number of weights in the neural network. Therefore, in each of the t training epochs we have to push n training examples through a neural network with n weights, hence the n^2 term. In contrast, the quantum approach would require an effort scaling as $n + t \cdot \log(n)$. For datasets where n reaches into the trillions, the resultant savings in time and resources could be monumental. While the realization of large-scale QRAM is still aspirational, the cost and efficiency benefits it promises would ultimately favor quantum methodologies. One might envision a future when vast repositories of data and human knowledge are compactly encoded into quantum states, empowering quantum machine-learning algorithms to train models using a fraction of the current energy and hardware costs. We should note that it is still an open research question how QRAM that allows neural networks to perform a query in $\log(n)$ time steps can be engineered in a fault-tolerant manner at a computational cost.²⁹

There are additional quantum advantages relevant to the regime of large-scale computation and big-data tasks underpinning modern AI that do not require QRAM. Encoding classical data in quantum states can lead to exponential savings in communication for problems like inference with distributed graph neural networks and gradient-based training.³⁰ For some natural graph problems, quantum algorithms provide exponential memory savings over any classical algorithm.³¹ Specifically, such discrete optimization problems and associated results in quantum communication complexity, which provide exponential savings in storage, suggest that more non-QRAM examples of quantum advantage exist for big data tasks. In many AI problems, memory, not computation, is the bottleneck. These types of advantages leverage the ability to represent information in quantum states that is in some sense a form of exponential compression yet crucially can still be used for computation. They make full use of the physical substrate of storage or communication that is allowed by quantum mechanics and that classical algorithms cannot take advantage of. A recent example is “quantum oracle sketching,” which ingests massive sequences of classical data by mapping them onto small rotations of a quantum state. This state can then be used for classification or dimensionality reduction, offering exponential advantages over classical systems that lack the memory capacity to process such vast data streams.³² Whenever computers communicate optically, they do not leverage the fact that one could instead send a quantum superposition of different configurations of light over the same communication channel, which can be used, for example, to solve classification problems that would otherwise require exponentially larger bandwidth. Large language models such as Google’s Gemini require multiple distant data centers to communicate during training, and thus communication bottlenecks are a constraint on such computation that could be alleviated with the help of quantum computers.³³

Recently, the term “agentic AI” has gained prominence, describing systems capable of effecting actions in the real world, such as booking a flight. This technical usage, however, is not intended to imply that artificial intelligence has agency or free will in a philosophical sense. Indeed, it is difficult to reconcile the concept of free will with a hardware substrate governed by the deterministic laws of classical mechanics and electrodynamics. Within such a framework, genuine choice and the emergence of agency – free will – become impossible to conceptualize. Everything proceeds on rails, and there is never truly a choice, a moment when agency could manifest. This does not fundamentally change when we add a source of randomness; if the decision between two alternatives is decided by the throw of a digital die, true agency remains absent.

Similarly, the computer science underlying today’s AI has yet to provide a widely accepted explanation for what implements human consciousness. Humans have feelings and enjoy experiences; many of us would argue that this is what life is all about. Likewise, many of us have a sense of agency, believing we are at least partially in control of our destiny. Yet artificial intelligence is entirely stumped when challenged with explaining these experiential features associated with biological intelligence, a dilemma philosopher David Chalmers dubbed the “hard problem.” To accurately predict the future of AI, we cannot dismiss these profound aspects of the human mind as mere epiphenomena and leave the discussion solely to philosophers.

Addressing these challenges, a recent article in *Entropy* coauthored by one of us puts forth a proposal drawing directly from the ontology of quantum mechanics, particularly aligning with the many-worlds formulation.³⁴ It suggests that consciousness is how we experience the emergence of a definite, classical world from the multitude of coexisting realities that, according to the many-worlds interpretation, the equations of quantum mechanics describe at any given moment. More specifically, the article suggests that a moment of consciousness arises whenever a quantum mechanical state enters into a superposition of two or more configurations. This moment of superposition, the authors argue in the article, is precisely the moment of choice when agency can manifest.

To our knowledge, this is the first proposal for the physical substance of consciousness that is amenable to experimental testing. This challenging but, in principle, feasible test consists of coherently linking quantum degrees of freedom in a human brain to a quantum computer, with the aim of investigating whether such a link allows for a controlled expansion of human conscious experience. To study how such a coherent link between qubits and neural tissue could be established, the article proposes experiments with xenon isotopes on brain organoids and fruit flies. Intriguingly, early (still unpublished) experimental data show that different xenon isotopes exhibit different anesthetic potencies. If this isotope effect is confirmed, then it would provide strong evidence that quantum processes need

to be considered when modeling the action of anesthetics or more generally when formulating a theory of consciousness.

Quantum generalizations have been proposed for most classical machine-learning methods. Take any acronym for a popular machine-learning (ML) method, put a Q in front of it, and a Google search will point you to a paper describing the quantum version (QML). Often such generalizations come with a theoretical argument about why they should be superior. For instance, quantum kernel methods can access large feature spaces inaccessible to their classical counterparts, quantum GANs (Generative Adversarial Networks) or quantum diffusion networks can work with probability distributions that can't be represented on classical computers. But do such theoretical advantages really matter in machine-learning practice? We don't know yet. This is because machine learning is manifestly a heuristic discipline. Whether a task is learnable or whether the architecture of a neural network is well suited for a learning task is typically found by trial and error. Indeed, classical neural networks were discovered experimentally, and it remains a major open research problem to analytically and rigorously explain why they are so effective. Similarly, it remains to be empirically tested whether the quantum generalizations of machine learning deliver better results. But the testing of heuristic QML algorithms will have to wait until quantum computers reach the size at which they can handle datasets big enough to be of practical relevance. Table 1 explores why different QML algorithms may have an advantage.

While most QML algorithms today have only theoretical advantages, there exists an early empirical foundation built on smaller-scale quantum processors: early Nuclear Magnetic Resonance platforms provided experimental proof-of-principle demonstrations for classifying quantum states and testing elementary QML protocols, bridging the gap between theoretical algorithms and physical realization. As hardware scales, these early results provide confidence that the theoretical speed-ups will manifest empirically.

While quantum computing holds promise for advancing artificial intelligence in the future, AI is already accelerating quantum technology today. This synergy is likely to create a virtuous cycle, leading to the emergence of new capabilities. Where is AI boosting quantum computing and where might its impact be greatest?

Quantum error correction – the technology that ensures that quantum information does not leak into the environment and that all operations necessary to execute a quantum algorithm are executed with high precision – is crucial for scalable quantum computing, but the hardware overhead to achieve it remains high. Machine learning is being deployed to discover more efficient methods for encoding quantum information stably and for decoding the error syndromes reported by

Table 1
Advantages of Various Quantum Machine Learning Algorithms

Classical AI/ ML Task	Underlying Computation	Quantum Algorithm(s)	Potential Quantum Advantage	Challenges/ Caveats
Model training	Parameter optimization	QAOA, VQE, DQI, quan- tum Gibbs sampling, quantum streaming	Faster conver- gence, escape local minima, find local minima	Superquadratic speedups are only established for very specific optimization problems, less memory
Generative modeling	Sampling complex distributions	QBMs, QGANs, quan- tum noise in diffusion networks	Higher expres- sivity, better sample quality	Practical advantage in generating com- pelling images, videos, or other patterns not yet established
Classification	Kernel compu- tation	QSVM (quan- tum kernels)	Access intrac- table feature spaces	Demonstrating advantage on real data, data encoding
Linear regression	Matrix inver- sion $Ax = b$	HHL algo- rithm	Exponential speedup in N (dimension)	Data I/O over- head, sensitivity to properties of the Matrix A
Dimensionality reduction	Eigenvalue decomposition	QPCA (quan- tum phase estimation)	Exponential speedup in N (dimension)	Data I/O over- head
Data search	Unstructured search	Grover's algo- rithm	Quadratic speedup	Oracle implementation efficiency, qua- dratic speedup requires long runtimes to manifest

Classical AI/ ML Task	Underlying Computation	Quantum Algorithm(s)	Potential Quantum Advantage	Challenges/ Caveats
Pattern matching	Finding hidden shifts	Montanaro, Kuperberg	Superpolyno- mial speedup	Unclear wheth- er this is a natu- rally occurring problem
Anomaly detection	Search for rare events	Grover's algorithm	Quadratic speedup	Oracle im- plementation efficiency, qua- dratic speedup requires long runtimes to manifest

Table 1 shows how quantum machine-learning algorithms have been suggested for most classical machine-learning tasks. The quantum algorithms abbreviated above are QAOA (Quantum Approximate Optimization Algorithm), VQE (Variational Quantum Eigensolver), DQI (Decoded Quantum Interferometry), QBMs (Quantum Boltzmann Machines), QGANs (Quantum Generative Adversarial Networks), QSVM (Quantum Support Vector Machine), HHL (Harrow–Hassidim–Lloyd), and QPCA (Quantum Principal Component Analysis). Source: Authors’ data and analysis.

qubit measurements designed to detect errors. A notable early success in this area is AlphaQubit, developed through a collaboration between the Google DeepMind and Quantum AI teams. This ML-based decoding system maps error syndromes to suspected errors. This task is well suited for transformer networks, the neural network architecture underlying large language models, because we can generate abundant synthetic training data by injecting errors into simulated circuits and measuring the resulting syndromes, which then enables us to fine-tune the network with experimental training data from actual hardware runs. The result was that AlphaQubit outperformed the best human-designed methods in error rate.³⁵

AI’s recent advancements in generating and refining programs for complex problems – exemplified by DeepMind’s AlphaEvolve tackling classical optimization challenges – suggest a powerful synergy with quantum computing.³⁶ By expanding the search space to include quantum programs, the AlphaEvolve coding agent could be employed to discover efficient quantum circuits for classically intractable problems.

Beyond software and algorithms, AI’s potential extends to the processes of designing and refining the quantum hardware itself, further fueling the virtuous

cycle. For instance, in chip design, machine-learning models can be trained on extensive databases of existing designs and their corresponding performance metrics. The Google Quantum AI team is actively exploring this, leveraging metrology data to teach AI how design parameters map to chip performance, with the goal of eventually enabling automated and optimized chip design based on a sufficiently large and comprehensive dataset.

Beyond the initial design phase, processor calibration presents another challenge for which artificial intelligence excels. Tuning a novel quantum processor to its optimal operational regime involves navigating a high-dimensional parameter space – a task akin to finely tuning a complex musical instrument. Here, sophisticated AI algorithms are indispensable for efficiently solving these intricate optimization problems and achieving peak performance from the quantum hardware. Furthermore, maintaining the quality and integrity of quantum hardware throughout the manufacturing process significantly benefits from AI-driven wafer inspection. Computer vision techniques, powered by machine learning, automate and enhance the classification of wafer and circuit integrity. This application is yet another example of how AI accelerates a critical step in the hardware development pipeline.

The fundamental ways by which quantum computing and artificial intelligence can help each other are hardware-agnostic and apply to all scalable approaches to quantum computing, including those based on superconducting qubits, neutral atoms, trapped ions, photons, spin qubits, or topological qubits.

To predict how quantum computing and AI will mutually enhance and evolve toward Quantum AI, it's instructive to recall AI's history (Figure 2): Following initial conceptual breakthroughs, significant progress has largely been driven by the scaling of underlying hardware. This scaling enables the creation of larger models and the ingestion of the vast datasets essential for their training. The perceptron, the first multilayer neural network and often considered the precursor to modern neural networks, was proposed by psychologist Frank Rosenblatt in 1957. The architecture of a modern deep neural network powering a large language model is not fundamentally different. Both consist of layers of McCulloch-Pitts neurons – a simplified computational model of a biological neuron that receives inputs, multiplies each by a corresponding weight, sums these weighted inputs, and generates an output if this sum exceeds a predefined threshold. A deep neural network is trained with the backpropagation algorithm, a development that started in the 1970s. However, it wasn't until the mid-1980s, particularly through the joint work of psychologist David Rumelhart and computer scientists Geoffrey Hinton and Ronald J. Williams, that its significance for training artificial neural networks was fully realized, becoming a cornerstone of the field. The reason we are seeing an explosion in the performance and use of neural networks is that the underlying classical hardware has become exponentially more performant. Moore's law once

Figure 2
Historic Milestones in Quantum Computing and AI

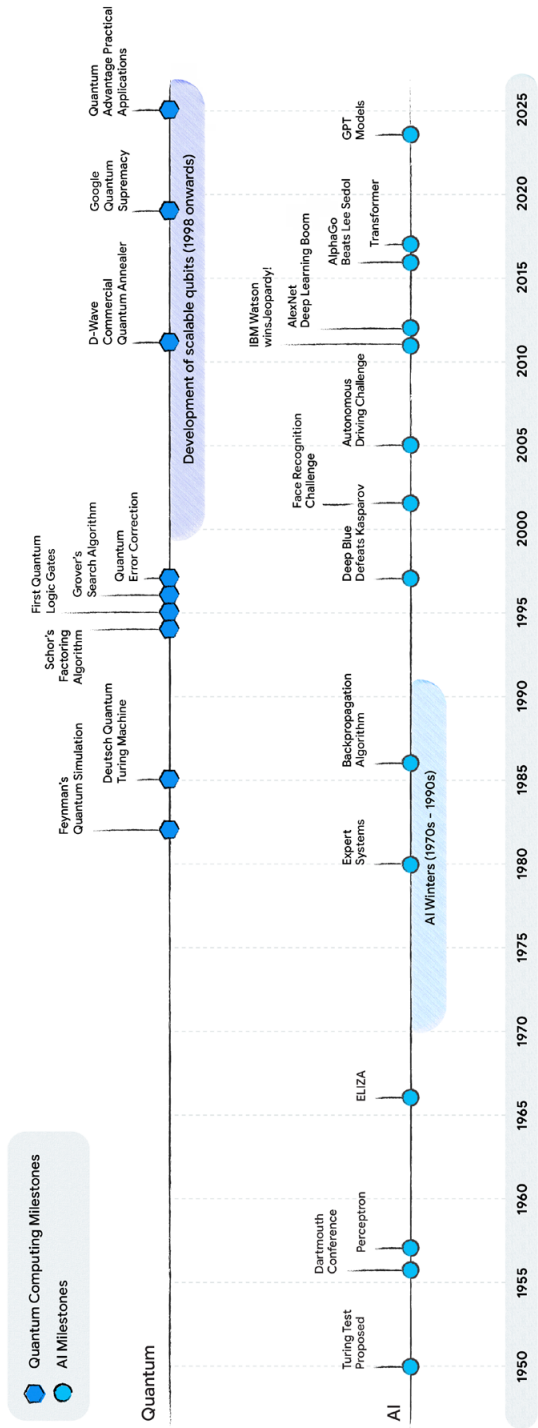


Figure 2 shows seminal events in the histories of quantum computing and AI. AI is about seventy-five years in the making; quantum computing is only about half as old. Our prediction is that well before AI turns one hundred, it will merge with quantum computing and morph into Quantum AI. Source : Figure by the authors.

ensured that the power of an individual processor doubled about every two years. As that progress began to slow, classical parallel processing compensated, leading to the current era in which massive data centers are filled with processors, such as graphics processing units (GPUs) and tensor processing units (TPUs), specialized for the tensor computations needed in neural network training and operation. This immense compute power, coupled with vast multimodal datasets, unlocked the potential of long-established AI concepts. Computational innovations such as the attention mechanism of transformer networks that allowed parallelization of the training for language models were certainly important, but the fundamental driver enabling the AI revolution has been hardware progress and large datasets. The machine-learning algorithms have remained remarkably unchanged.

Now the hardware powering artificial intelligence is once again running up against limitations. Besides the costs of so many TPUs or GPUs, the energy need is becoming substantial. Generating a single image uses about 10 watt-hours, roughly the amount of energy to charge a cell phone; training a large language model requires about 10 gigawatt-hours, an amount a nuclear power plant outputs in ten hours. Can we expect that quantum hardware will help to alleviate these hurdles and to ensure the continued progress of AI? Because of the new capabilities afforded by quantum computers discussed in this essay, the answer seems to be “yes.” The fact that vast amounts of training data can be stored and manipulated succinctly in superposition is a key reason why quantum resources may one day curb the steeply rising costs of training large neural networks.

But beyond these practical efficiencies, there’s a more fundamental argument: Since information must be represented by configurations of matter, it is ultimately the laws of physics that dictate which operations are possible within any information-processing system. It follows that the most advanced AI systems will eventually leverage the richest computational framework allowed by physics, perhaps eventually even incorporating principles from yet undiscovered physics in quantum gravity and elsewhere, should such theories yield more powerful computational models. From this perspective, continued progress in AI seems to lead inexorably toward Quantum AI. Hence, we dare to predict that by the time artificial intelligence turns one hundred years old, it will have morphed into Quantum AI.

ABOUT THE AUTHORS

Maria Spiropulu is a particle physicist and the Shang-Yi Ch'en Professor of Physics at the California Institute of Technology. She is Co-founder of Bohr Quantum Technology, an advanced quantum networking technologies startup. Her recent research includes testing wormhole teleportation protocols, advancing quantum teleportation networks, optimizing real time AI on field-programmable gate arrays and application-specific integrated circuits, and developing and building an ultraprecision multichannel timing sensor system for the High-Luminosity Large Hadron Collider at CERN. She was Cochair of *Elementary Particle Physics: The Higgs and Beyond 2025*, a consensus study report for the National Academies of Sciences, Engineering, and Medicine that presents a forty-year strategic vision for the field of particle physics. She also coauthored the Department of Energy's 2020 *Quantum Internet Blueprint* report.

Hartmut Neven is Vice President of Engineering at Google, where he leads the Quantum Artificial Intelligence Lab, which he founded in 2012. The Google Quantum AI team has achieved a number of firsts, including implementing the first quantum computations beyond the reach of classical supercomputers, demonstrating quantum error correction, and executing the first verifiable quantum computation.

ENDNOTES

- ¹ Wikipedia Contributors, "APX," Wikipedia, last edited March 24, 2025, <https://en.wikipedia.org/wiki/APX>.
- ² Stephen P. Jordan, Noah Shutty, Mary Wootters, et al., "Optimization by Decoded Quantum Interferometry," arXiv (2024), <https://arxiv.org/abs/2408.08292>; and Edward Farhi, Sam Gutmann, Daniel Ranard, and Benjamin Villalonga, "Lower Bounding the MaxCut of High Girth 3-Regular Graphs Using the QAOA," arXiv (2025), <https://doi.org/10.48550/arXiv.2503.12789>.
- ³ Stephen P. Jordan and Google Quantum AI, "Quantum Algorithm Zoo," last updated October 30, 2025, <https://quantumalgorithmzoo.org>.
- ⁴ "XPRIZE Quantum Applications," XPRIZE Foundation, <https://www.xprize.org/prizes/qc-apps> (accessed May 2, 2025).
- ⁵ Stefan Chmiela, Valentin Vassilev-Galindo, Oliver T. Unke, et al., "Accurate Global Machine Learning Force Fields for Molecules with Hundreds of Atoms," *Science Advances* 9 (2) (2023), <https://doi.org/10.1126/sciadv.adf0873>.
- ⁶ Ibid.; and James Kirkpatrick, Brendan McMorro, David H. P. Turban, et al., "Pushing the Frontiers of Density Functionals by Solving the Fractional Electron Problem," *Science* 374 (6573) (2021): 1385–1389.
- ⁷ John Jumper, Richard Evans, Alexander Pritzel, et al., "Highly Accurate Protein Structure Prediction with AlphaFold," *Nature* 596 (7873) (2021): 583–589; and Kirkpatrick, McMorro, Turban, et al., "Pushing the Frontiers of Density Functionals by Solving the Fractional Electron Problem."
- ⁸ Dmitry A. Abanin, Rajeev Acharya, Laleh Aghababaie-Beni, et al., "Constructive Interference at the Edge of Quantum Ergodic Dynamics," arXiv (2025), <https://doi.org/10.48550/arXiv.2506.10191>; and C. Zhang, R. G. Cortiñas, A. H. Karamlou, et al., "Quantum

- Computation of Molecular Geometry via Many-Body Nuclear Spin Echoes,” arXiv (2025), <https://doi.org/10.48550/arXiv.2510.19550>.
- ⁹ Nicholas C. Rubin, Dominic W. Berry, Alina Kononov, et al., “Quantum Computation of Stopping Power for Inertial Fusion Target Design,” *Proceedings of the National Academy of Sciences* 121 (23) (2024), <https://doi.org/10.1073/pnas.231772121>.
- ¹⁰ Robin Kothari and Rolando Somma, “A New Quantum Algorithm for Classical Mechanics with an Exponential Speedup,” The Google Research Blog, December 4, 2023, <https://research.google/blog/a-new-quantum-algorithm-for-classical-mechanics-with-an-exponential-speedup>; and Aram W. Harrow, Avinatan Hassidim, and Seth Lloyd, “Quantum Algorithm for Linear Systems of Equations,” *Physical Review Letters* 103 (15) (2009): 150502.
- ¹¹ Juan Maldacena and Leonard Susskind, “Cool Horizons for Entangled Black Holes,” *Fortschritte der Physik* 61 (9) (2013): 781–811, <https://doi.org/10.1002/prop.201300020>; and Ping Gao, Daniel Louis Jafferis, and Aron C. Wall, “Traversable Wormholes via a Double Trace Deformation,” *Journal of High Energy Physics* 2017 (12) (2017): 1–25.
- ¹² Daniel Jafferis, Alexander Zlokapa, Joseph D. Lykken, et al., “Traversable Wormhole Dynamics on a Quantum Processor,” *Nature* 612 (7938) (2022): 51–55.
- ¹³ Juan Maldacena and Douglas Stanford, “Remarks on the Sachdev-Ye-Kitaev Model,” *Physical Review D* 94 (10) (2016): 106002.
- ¹⁴ Garnet Kin-Lic Chan and Sandeep Sharma, “The Density Matrix Renormalization Group in Quantum Chemistry,” *Annual Review of Physical Chemistry* 62 (2011): 465–481.
- ¹⁵ Srihari S. Iyengar, “Quantum Chemical Dynamics at the Nanoscale: The Role of Nuclear Quantum Effects, Non-Adiabaticity, and Time-Dependence,” *Chemical Reviews* 123 (1) (2023): 405–466.
- ¹⁶ Hsin-Yuan Huang, Michael Broughton, Jordan Cotler, et al., “Quantum Advantage in Learning from Experiments,” *Science* 376 (6598) (2022): 1182–1186.
- ¹⁷ Google Research, “Quantum Transduction and Networking for Scalable Computing Applications,” <https://research.google/programs-and-events/quantum-transduction-and-networking-for-scalable-computing-applications> (accessed May 3, 2025).
- ¹⁸ Richard R. Allen, Francisco Machado, Isaac L. Chuang, et al., “Quantum Computing Enhanced Sensing,” arXiv (2025), <https://doi.org/10.48550/arXiv.2501.07625>.
- ¹⁹ Maldacena and Stanford, “Remarks on the Sachdev-Ye-Kitaev Model.”
- ²⁰ Akash V. Dixit, Srivatsan Chakram, Kevin He, et al., “Searching for Dark Matter with a Superconducting Qubit,” *Physical Review Letters* 126 (14) (2021): 141302.
- ²¹ Aaron Chou, Kent Irwin, Reina H. Maruyama, et al., “Quantum Sensors for High Energy Physics,” arXiv (2023), <https://doi.org/10.48550/arXiv.2311.01930>.
- ²² Feng-Hsiung Hsu, *Behind Deep Blue: Building the Computer That Defeated the World Chess Champion* (Princeton University Press, 2022).
- ²³ Volodymyr Mnih, Koray Kavukcuoglu, David Silver, et al., “Human-Level Control Through Deep Reinforcement Learning,” *Nature* 518 (7540) (2015): 529–533.
- ²⁴ David Silver, Aja Huang, Chris J. Maddison, et al., “Mastering the Game of Go with Deep Neural Networks and Tree Search,” *Nature* 529 (7587) (2016): 484–489.
- ²⁵ Scott Aaronson, *Quantum Computing Since Democritus* (Cambridge University Press, 2013).

- ²⁶ Sadiq Muhammad, Armin Tavakoli, Maciej Kurant, et al., “Quantum Bidding in Bridge,” *Physical Review X* 4 (2) (2014): 021047; Joseph X. Lin, Joseph A. Formaggio, Aram W. Harrow, and Anand V. Natarajan, “Quantum Blackjack or Can MIT Bring Down the House Again?” arXiv (2019), <https://arxiv.org/abs/1908.09417>; Dawei Ding and Liang Jiang, “Coordinating Decisions via Quantum Telepathy,” arXiv (2024), <https://doi.org/10.48550/arXiv.2407.21723>; and Dar Gilboa, Hagay Michaeli, Daniel Soudry, and Jarrod R. McClean, “Exponential Quantum Communication Advantage in Distributed Inference and Learning,” in *NIPS ’24: Proceedings of the 38th International Conference on Neural Information Processing Systems*, ed. Amir Globerson, Lester Mackey, Danielle Belgrave, et al. (Curran Associates, Inc., 2024), 30425–30473.
- ²⁷ Richard Sutton, “The Bitter Lesson,” March 13, 2019, <http://www.incompleteideas.net/IncIdeas/BitterLesson.html>; and Jared Kaplan, Sam McCandlish, Tom Henighan, et al., “Scaling Laws for Neural Language Models,” arXiv (2020), <https://doi.org/10.48550/arXiv.2001.08361>.
- ²⁸ Alexander Zlokapa, Hartmut Neven, and Seth Lloyd, “A Quantum Algorithm for Training Wide and Deep Classical Neural Networks,” arXiv (2021), <https://doi.org/10.48550/arXiv.2107.09200>.
- ²⁹ Vittorio Giovannetti, Seth Lloyd, and Lorenzo Maccone, “Quantum Random Access Memory,” *Physical Review Letters* 100 (16) (2008): 160501; and Alexander M. Dalzell, András Gilyén, Connor T. Hann, et al., “A Distillation-Teleportation Protocol for Fault-Tolerant QRAM,” arXiv (2025), <https://doi.org/10.48550/arXiv.2505.20265>.
- ³⁰ Gilboa, Michaeli, Soudry, and McClean, “Exponential Quantum Communication Advantage in Distributed Inference and Learning.”
- ³¹ Chi-Fang Chen, Hsin-Yuan Huang, John Preskill, and Leo Zhou, “Local Minima in Quantum Systems,” in *STOC 2024: Proceedings of the 56th Annual ACM Symposium on Theory of Computing* (Association for Computing Machinery, 2024), 1323–1330.
- ³² Haimeng Zhao, Alexander Zlokapa, Hartmut Neven, et al., “Exponential Quantum Advantage in Processing Massive Classical Data,” arXiv (2026), <https://arxiv.org/pdf/2604.07639>.
- ³³ John Kallaugher, Ojas Parekh, and Nadezhda Voronova, “Exponential Quantum Space Advantage for Approximating Maximum Directed Cut in the Streaming Model,” in *STOC 2024: Proceedings of the 56th Annual ACM Symposium on Theory of Computing* (Association for Computing Machinery, 2024), 1805–1815; Seid Koudia, Angela Sara Cacciapuoti, Kyrylo Simonov, and Marcello Caleffi, “How Deep the Theory of Quantum Communications Goes: Superadditivity, Superactivation and Causal Activation,” *IEEE Communications Surveys & Tutorials* 24 (4) (2022): 1926–1956; and Gemini Team Google, “Gemini: A Family of Highly Capable Multimodal Models,” arXiv (2023), <https://doi.org/10.48550/arXiv.2312.11805>.
- ³⁴ Hartmut Neven, Adam Zalcman, Peter Read, et al., “Testing the Conjecture That Quantum Processes Create Conscious Experience,” *Entropy* 26 (6) (2024): 460.
- ³⁵ Johannes Bausch, Andrew W. Senior, Francisco J. H. Heras, et al., “Learning High-Accuracy Error Decoding for Quantum Processors,” *Nature* 635 (8040) (2024): 834–840, <https://doi.org/10.1038/s41586-024-08148-8>.
- ³⁶ Bernardino Romera-Paredes, Mohammadamin Barekatin, Alexander Novikov, et al., “Mathematical Discoveries from Program Search with Large Language Models,” *Nature* 625 (7995) (2024): 468–475.

The Future of AI-Facilitated Medicine

Eric J. Topol

AI's most significant contribution to medicine to date is its ability to produce more accurate and comprehensive interpretations of medical images, as validated through randomized clinical trials. Next up is the opportunity to address the steady erosion of the patient-doctor relationship over several decades and a global burnout crisis among clinicians. By reducing the data clerical work of clinicians and giving patients more autonomy, AI has the potential to restore the humanity in medicine and care. Freeing physicians from constant screen time could bring back the physician's presence during clinic visits and help foster empathy and trust. Furthermore, AI's ability to aggregate and contextualize a "full stack" of patient data can enable new opportunities for prevention of major age-related diseases, significantly extending human health span. Despite this vast potential, the implementation of medical AI faces substantial challenges, including mitigating bias, ensuring data privacy and security, and establishing appropriate regulatory and reimbursement frameworks.

The pioneers and leaders in the field of artificial intelligence have consistently singled out advances in medicine and health care as its most substantial and unequivocally positive impact. As Geoffrey Hinton put it: "I always pivot to medicine as an example of all the good it can do because almost everything it's going to do there is going to be good."¹ Similarly, speaking about the end of disease, Demis Hassabis said: "I think that's within reach. Maybe within the next decade or so, I don't see why not."² Then there's Dario Amodei, who believes AI will cure most cancers, prevent Alzheimer's, and double the human lifespan, and Sam Altman, who predicted that "we will see diseases cured at an unprecedented rate thanks to AI."³ And there's Mustafa Suleyman, who said: "One of the applications that I'm most excited about is this quest for medical superintelligence. I think it's pretty clear in the next couple of years we are going to build models that are better than the vast majority of expert clinicians in any discipline at diagnosis."⁴

Encouraging as these proclamations are, the specifics of how AI can and will ultimately transform medicine remain vague. This essay intends to address the gaps and provide a futuristic perspective of how AI can achieve profound improvements in the practice of medicine and in health outcomes for patients – by improving diagnostic accuracy, restoring humanity to the doctor-patient relationship, and preventing diseases.

The medical community tends to minimize its errors. However, an important report by researchers at Johns Hopkins estimates that nearly eight hundred thousand Americans per year suffer permanent disability or death due to significant diagnostic medical errors.⁵ And that doesn't include errors in treatment, which represent another serious problem. With the common use of the term "precision medicine," there is no acknowledgment that if you make the same mistake repeatedly, it is, in a perverse way, very precise. We desperately need to promote diagnostic accuracy in medicine.

Today we're making strides, particularly with respect to interpretation of medical imaging. The largest randomized trial in medical AI compared radiologists using AI against radiologists without AI in a sample of more than one hundred thousand women undergoing mammography.⁶ The radiologists working with AI support detected approximately 30 percent more instances of breast cancer and 25 percent more invasive tumors than radiologists without AI. Similarly, across forty randomized trials in gastroenterology, machine vision–supported colonoscopy (compared with colonoscopy without machine vision) increased the overall detection of adenomas and polyps by more than 20 percent.⁷

Beyond these randomized trials, there is supportive evidence for improved accuracy of all types of medical images, including X-rays, ultrasound, positron emission tomography (PET) scans, CT scans, and MRI. The next step is for the initial image review process to be fully automated, with AI systems capable of not just screening specific traits but of synthesizing information to produce comprehensive reports that can be reviewed by supervising physicians. Researchers recently did just that in one of the most complex areas of medical imaging: echocardiography. The motion of the heart and the need to acquire and interpret multiple views of not just static images but of ultrasound video create a data-rich and complex environment in which to develop a foundation AI model. Trained using over twelve million echocardiographic videos and paired text reports across twenty-three benchmarks of heart structure and function, the EchoPrime model achieved state-of-the-art performance in interpretation of cardiac features and diagnoses.⁸ There is work in progress to do the same for all medical scans, with the potential to improve diagnosis accuracy and streamline workflows.

That brings us to the concept of *superhuman vision*. The conventional interpretation of medical images – answering a specific query – may become obsolete. For example: Does a chest X-ray show pneumonia? With computer vision, we now can detect information embedded in chest X-rays about the presence of type 2 diabetes, the risk of atherosclerosis in the next ten years, the coronary artery calcium score, ejection fraction, and osteoporosis. The same is true for mammograms that can accurately detect heart disease risk, or chest CT scans finding pancreatic cancer. This "opportunistic" medical scan interpretation is a reflection of the power of digital eyes to "see" things that expert clinicians cannot.⁹

Superhuman vision is not limited to radiologic scans. Perhaps the most extraordinary example is the retina image, whether it be a photo or optical coherence tomography (OCT). Various studies, and more recently AI foundation models, which use vast and diverse datasets for a wide range of tasks, have indicated how remarkably informative the retina can be.¹⁰ AI interpretation of retinal imaging can tell us about a patient's blood pressure, glucose control, and risk of heart, kidney, liver, and gall bladder disease. The images are a window into the future risk of Alzheimer's and Parkinson's disease in people without any symptoms. In the retina, AI can also detect the likelihood of subclinical (no symptoms, no narrowings that limit blood flow) atherosclerosis of the coronary arteries and silent brain injury, as well as the risk of stroke.

The aforementioned 12-lead electrocardiogram likewise lends itself to AI insights that were not previously obtainable through imaging alone. They include the determination of the age and sex of the patient; the presence of anemia, diabetes, or kidney disease; ejection fraction of the left ventricle; filling pressure of the heart; prediction of atrial fibrillation; and the risk of stroke. Or, with AI, a pathology slide of a questionable cancer tissue can determine the genomic driver mutations of the tumor, its malignancy status, the site of origin, and patient prognosis. The possibilities of what all is encoded in different types of medical images seem limitless, and were certainly not envisioned before the advent of deep learning and generative AI. Whether and when all this information will move from research publications to medical practice is uncertain, and will require further validation, but it is fair to say we won't be leaving so much data on the table thanks to the power of AI superhuman vision.

AI is making great progress toward more accurate and comprehensive diagnostic imaging, but the contributions of AI for promoting accuracy go well beyond that. Specially adapted large language models have demonstrated the ability to concisely and accurately synthesize all of a patient's data, which can be quite extensive and from multiple sources.¹¹ Improved AI performance extends to better patient communication and management plans. In one study involving 150 case scenarios and patient actors, AI demonstrated superior performance to specialist clinicians in thirty of thirty-two tasks.¹² While that study is not representative of the real world of medical practice, there has since been a randomized trial in real-world medical centers from four continents that confirms the advantages of using AI for ophthalmology care with respect to diagnosis of seven major eye diseases and patient management.¹³ The clinical benefits of AI are also not limited to high-income countries: A randomized trial of primary care at fifteen clinics in Nairobi involving nearly forty thousand patient visits showed that clinician use of an AI consult tool reduced diagnosis and treatment errors. Similar results have followed for many other medical tasks, such as diagnosis of tuberculosis, childhood cancer, diabetic retinopathy, and rheumatic heart disease in low- and middle-income countries.¹⁴

For several decades, there has been a steady erosion of the patient-doctor relationship, in part because of the very limited time allotted to clinic visits and bedside hospital rounds. The business of medicine and the drive for clinician productivity have left physicians with the sense they cannot provide adequate care for patients – the principal reason they entered the profession. A typical doctor-patient encounter today lasts only a few minutes without much, if any, eye-to-eye contact – the legal and ethical requirements of progress notes push doctors from the bedside to the keyboard as they fulfill data-clerk responsibilities. The patients perceive the doctor’s lack of presence, which in turn engenders a lack of trust, care, and concern. What was an intimate human-human bond decades ago has slipped to an estranged encounter, devoid of compassion, often with very limited physical examination and laying of hands. In my 2019 book *Deep Medicine*, I wrote that there was potential for AI to foster humanity in medicine.¹⁵ Many find that idea to be counterintuitive: How could technology make medicine more human again?

Liberation from the keyboard could markedly improve physicians’ visits with a patient. With the use of AI scribes that synthesize notes with many downstream functions, doctors can redirect their attention to the patient, bringing back face-to-face communication. At the same time, ambient conversations can produce excellent notes with audio links for the patient to return to after the appointment. On top of that, the AI assistant can schedule labs, follow-up appointments, and scans; run preauthorizations with insurance companies and submit prescriptions; do coding and billing; and even nudge patients after the visit to see if they are following up on planned actions, such as measuring their blood pressure or going for walks. In this way, administrative use of AI is humanistic. The automation is a major step toward relieving clinicians of their data-clerk burden and allowing them to focus more fully on their patients. Such systems have already been shown to save clinician time and reduce burnout, a global crisis among health care workers.¹⁶ The uptake of this technology has been remarkably fast; before long, automated note-taking and AI support for time-consuming administrative tasks will be the norm.

And patients are gaining empowerment. Even prior to generative AI, algorithmic diagnostics were changing how patients monitor their health, enabling them to measure abnormal heart rhythm by smartwatch, screen for diabetic retinopathy at grocery stores, detect urinary tract infections via an over-the-counter kit, and classify skin rashes or lesions by app.¹⁷ Now patients are using AI chatbots to make diagnoses (some of which had eluded doctors and workups), get first and second opinions, enter their lab data, obtain advice for optimal treatment, or prepare for a doctor visit.¹⁸ This trend was reflected in a recent *New York Times* article entitled “Empathetic, Available, Cheap: When AI Offers What Doctors Don’t.”¹⁹

I had not considered that empathy could be facilitated by machines. But in a systematic review of fifteen studies comparing patient perception of empathy, all but two of the studies showed that patients reported higher levels of empathy in

their text conversations with AI-based chatbots than in their text communications with clinicians.²⁰ While AI isn't empathetic, its ability to foster a sense of empathy through appropriate language is significant inasmuch as the perception of empathy promotes better patient outcomes. In the future, this ability will likely lead to AI coaching of clinicians to be better, more empathetic communicators. Going further, think of how AI review of the physician's automated note could lead to productive questioning: "Why did you interrupt Mrs. Jones after nine seconds? Why didn't you give her the chance to express her real concerns?" While this was certainly not envisioned as a role for AI and many clinicians will bristle at the thought, it would not be surprising for such coaching to become part of an annual licensing requirement for the practice of medicine. For clinicians to actualize the benefits of improved AI accuracy, recognize its drawbacks, and promote trust with patients, specific training with these new systems will be necessary.

Improving the patient-doctor relationship requires the gift of time; it is the starting point for deeper physician-patient discussion and careful physical examination. But AI also introduces the potential for health care administrators to use the technology to enhance productivity and profitability: instead of creating more time for meaningful doctor-patient interactions, these tools could simply allow clinicians to see more patients in less time. Such an approach to these technologies would be destructive and obviate the chance for AI to help restore the essential humanity in health care. To counter this risk, I have called for solidarity among clinicians to stand up to administrators.²¹ Clinicians should use their bargaining power to increase the time allotted for in-person outpatient visits, reserving this precious encounter for important matters while taking advantage of alternative patient services including AI chatbots, telehealth, and virtual coaches, when appropriate.²²

Cancer, cardiovascular disease, and neurodegenerative disorders are the three main age-related diseases. Cumulatively, and with onset usually after age sixty, they are the primary reasons why the human health span – the years lived in good health – is cut short. Besides their increasing incidence with age, they share two important features: 1) a long incubation period of approximately two decades before symptoms manifest and 2) an immune system dysregulation and loss of protection, with untoward inflammation. The terms "immunosenescence" and "inflammaging" are used to characterize the latter.

Primary prevention is a treatment approach to take preemptive steps to anticipate and address a person's risk and stop a disease before it occurs. Although frequently discussed, primary prevention of the three major age-related diseases is rarefied in medical practice. Mass cancer screening is done with mammography, colonoscopy, and prostate-specific antigen, but these tests, at best, are used for detection, not prevention of cancer. Once cancer is detected by a scan there are already billions of cancer cells present and an increased chance it has spread within the body. But

there are now ways through genomics to know who is at risk of cancer, including polygenic risk scores and whole genome sequencing. Further, our intact immune system is essential to prevent the occurrence of cancer and we have only recently found ways to determine the health and functionality of our immune responsiveness. Until now, surprisingly, there has not been any test in the clinic to gauge the health of our immune system. But the recent advance of organ clocks – a panel of blood proteins that tracks the pace of aging for each organ of the body and our immune system – offers a way forward.²³

While there has been a marked reduction of heart attacks in the past two decades, there are still more than eight hundred thousand per year in the United States that are not prevented. Focus has recently shifted to identifying vulnerable patients with inflamed atherosclerotic plaque, well before a potentially catastrophic event occurs.²⁴

For Alzheimer's disease, the breakthrough blood test known as p-tau217 can now detect the risk of this condition many years before any symptoms of mild cognitive impairment.²⁵ The risk of Alzheimer's is also increased in people with a family history, one or two copies of the APOE4 allele, and a high polygenic risk score. Brain biomarkers like p-tau217, similar to LDL cholesterol for heart disease, serve a dual purpose of defining risk and measuring changes from lifestyle modifications and medications.

With a long runway of being able to identify high-risk individuals for each of these age-related diseases, the opportunity to achieve primary prevention has never been more realistic and attainable. But it requires having the “full stack” of patients' data, including electronic health records, lifestyle factors, labs, images, retina scan, genomics, proteins (organ clocks), relevant blood biomarkers, environmental exposures, and wearable sensor data. We've learned that while doctors and patients typically dismiss lab tests in the normal range as unimportant, AI can pick up significant trends and contextualize these results with the medical records.²⁶ AI can also draw important data from the notes themselves, both structured and unstructured components, such as for detecting a high risk for pancreatic cancer many years before this difficult and deadly diagnosis could otherwise be made.²⁷ I have reviewed the opportunistic capabilities of medical scans and retinal imaging for predicting risk of cardiovascular and neurodegenerative diseases. In addition, AI models can now differentiate between “normal” mammograms, reclassifying some as showing a high risk of breast cancer in the next five years.²⁸

In aggregate, these layers of data set up multimodal AI to determine the risk status of each individual for one or more of the major age-related diseases and to track whether the risk has been decreased via interventions. The interventions include not only healthy lifestyle factors – such as an anti-inflammatory diet, regular exercise, and good sleep – but also drug treatments. The field of AI drug discovery has garnered much attention, with dozens of companies active in this space, some part-

nered with the major pharmaceutical firms. We tend to think of a drug as solely for treatment of symptoms and disease, but in the years ahead, there will be a shift to preventive medications. Perhaps the pluripotent action of glucagon-like peptide (GLP-1) family of drugs, with benefits that have extended from treating diabetes and obesity to preventing heart, kidney, liver, and addiction disorders (to some extent independent of weight loss), will be viewed as the turning point for preventive drugs. While optimizing lifestyle will always be regarded as fundamental and low-cost, it is clear more will be needed to achieve prevention of the big three age-related diseases.

Adding to newfound capabilities to predict disease is the publication of the first large health model, known as Delphi-2M.²⁹ Researchers built and validated Delphi-2M from electronic health records and lifestyle data of four hundred thousand UK Biobank participants and nearly two million people in Denmark. By tokening each individual's past data, the model, based on GPT-2, a relatively low-parameter and early generative AI precursor, was able to predict disease over the next twenty years – both the *disease or health event* and *when* – with a surprising level of accuracy. The prediction performance increased when polygenic risk scores were introduced as an additional layer of data. In the future, more advanced large health models will be validated with the full stack of data described above, taking the predictive capability to higher levels of accuracy.

There is also considerable potential for a digital-twin resource to predict diseases.³⁰ Previously, computer scientist and investor Kai-Fu Lee and I envisioned a global data resource with billions of individuals and their full stacks of medical and health data, along with all data captured during an extended follow-up period. With such a dataset – whose privacy and security must be protected by federated learning and encryption – we would have the capability of using nearest-neighbor analysis to find digital twins for any individual. Such data analysis would complement individuals' multimodal AI and large health models as another means of forecasting health events and disease risk to the highest level of accuracy, enabling primary prevention. No such resource yet exists, but the potential to promote human health span is too great to ignore, and it is likely an inevitability. The enormous economic upside of reducing the burden of these age-related diseases, concurrent with the marked aging of the U.S. population, makes investment in these AI tools and resources and the ideal of primary prevention exceptionally alluring.

While there is vast potential for AI to promote medical accuracy, humanism, and prevention in the future, there is no shortage of challenges. At this juncture, large language models lack readiness for broad medical use, with notable deficiencies determined by stress testing and representative medical benchmarks.³¹ I have mentioned the importance of data privacy and security, something we do not have a good track record with to date. But a shift to individuals

owning their own medical data on a secure platform, which could also incorporate data from sensors and genomics that are not currently part of our electronic health records, could lead to improvement. As for intrinsic cultural biases of these and other inputs, we must fully interrogate them to avoid their propagation. I have argued that AI can reduce health inequities, but the fact is that without deliberate initiatives that prioritize fairness, AI systems are likely to exacerbate them. Of course, AI-induced medical errors are also a concern, especially where there is direct connection between generative AI systems and patients with insufficient clinician oversight. We must lead a dedicated effort to interrogate and refine any patient-facing models before making them publicly available. To date, we have no objective data for how direct patient access is faring with respect to benefit and harm; that needs to be captured and monitored on an ongoing basis. There are also significant risks and issues with an approach of permissive regulatory oversight that has chiefly relied on low-bar clearance instead of formal approvals for applications of AI algorithms. Reimbursement standards for AI in health care markets have not adapted to the progress being made: stakeholder incentives remain misaligned, which contributes to stasis and delayed adoption of useful technologies. We must act on and build upon the recommendations we have for how to best proceed.³² And most important of all, we need to establish trust in the medical community and the public, which stems from transparency and compelling medical research to identify benefits and harms. So much of the medical AI research today is left compartmentalized, not leading to implementation in medical practice. That gaping chasm between research and practice needs to be bridged.

Medical AI's future impact will be transformative, especially if it attends to and eventually overcomes the many obstacles and challenges ahead. I believe that preventing age-related diseases, thereby significantly extending human health span, will be the single most important contribution of AI in medicine in the decades ahead.

ABOUT THE AUTHOR

Eric J. Topol is Chair of the Department of Translational Medicine at Scripps Research Institute, Director and Founder of the Scripps Research Translational Institute, Senior Consultant for the Division of Cardiovascular Diseases at the Scripps Clinic, and the Gary and Mary West Chair of Innovative Medicine in the Department of Translational Medicine at Scripps Research. He is the author of *Super Agers: An Evidence-Based Approach to Longevity* (2025), *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again* (2019), and *The Patient Will See You Now: The Future of Medicine is in Your Hands* (2014).

ENDNOTES

- ¹ Eric Topol, “Geoffrey Hinton : Large Language Models in Medicine. They Understand and Have Empathy,” *Ground Truths* (podcast), December 8, 2023, <https://erictopol.substack.com/p/geoffrey-hinton-large-language-models>.
- ² Scott Pelley, “Artificial Intelligence Could End Disease, Lead to ‘Radical Abundance,’ Google DeepMind CEO Demis Hassabis Says,” CBS News, August 3, 2025, <https://www.cbsnews.com/news/artificial-intelligence-google-deepmind-ceo-demis-hassabis-60-minutes-transcript>.
- ³ Anderson Cooper and Nichole Marks, “Why Anthropic CEO Dario Amodei Spends So Much Time Warning of AI’s Potential Dangers,” CBS News, November 16, 2025, <https://www.cbsnews.com/news/anthropic-ceo-dario-amodei-warning-of-ai-potential-dangers-60-minutes-transcript>; and Matteo Wong, “AI Executives Promise Cancer Cures. Here’s the Reality,” *The Atlantic*, April 25, 2025, <https://www.theatlantic.com/technology/archive/2025/04/how-ai-will-actually-contribute-cancer-cure/682607>.
- ⁴ Dan Fox and Richard Hodson, “The Future of AI,” *Nature*, November 14, 2025, <https://www.nature.com/immersive/d41586-025-03701-5/index.html#section-mToJSji4Ai>.
- ⁵ David E. Newman-Toker, Najlla Nassery, Adam C. Schaffer, et al., “Burden of Serious Harms from Diagnostic Error in the USA,” *BMJ Quality & Safety* 33 (2) (2024): 109–120, <https://doi.org/10.1136/bmjqs-2021-014130>.
- ⁶ Veronica Hernstrom, Viktoria Josefsson, Hanna Sartor, et al., “Screening Performance and Characteristics of Breast Cancer Detected in the Mammography Screening with Artificial Intelligence Trial (Masai): A Randomised, Controlled, Parallel-Group, Non-Inferiority, Single-Blinded, Screening Accuracy Study,” *Lancet Digital Health* 7 (3) (2025): [https://doi.org/10.1016/S2589-7500\(24\)00267-X](https://doi.org/10.1016/S2589-7500(24)00267-X).
- ⁷ Saeed Soleymanjahi, Jack Huebner, Lina Elmansy, et al., “Artificial Intelligence-Assisted Colonoscopy for Polyp Detection: A Systematic Review and Meta-Analysis,” *Annals of Internal Medicine* 177 (12) (2024): 1652–1663, <https://doi.org/10.7326/ANNALS-24-00981>.
- ⁸ Milos Vukadinovic, I-Min Chiu, Xiu Tang, et al., “Comprehensive Echocardiogram Evaluation with View Primed Vision Language AI,” *Nature* 650 (8103) (2026), <https://doi.org/10.1038/s41586-025-09850-x>.
- ⁹ Eric J. Topol, “AI-Enabled Opportunistic Medical Scan Interpretation,” *Lancet* 403 (10439) (2024): 1842, [https://doi.org/10.1016/S0140-6736\(24\)00924-3](https://doi.org/10.1016/S0140-6736(24)00924-3).
- ¹⁰ Eric J. Topol, “As Artificial Intelligence Goes Multimodal, Medical Applications Multiply,” *Science* 381 (6663) (2023): adk6139, <https://doi.org/10.1126/science.adk6139>.
- ¹¹ Dave Van Veen, Cara Van Uden, Louis Blankemeier, et al., “Adapted Large Language Models Can Outperform Medical Experts in Clinical Text Summarization,” *Nature Medicine* 30 (4) (2024): 1134–1142, <https://doi.org/10.1038/s41591-024-02855-5>.
- ¹² Tao Tu, Mike Schaeckermann, Anil Palepu, et al., “Towards Conversational Diagnostic Artificial Intelligence,” *Nature* 642 (8067) (2025): 442–450, <https://doi.org/10.1038/s41586-025-08866-7>.
- ¹³ Yilan Wu, Bo Qian, Tingyao Li, et al., “An Eyecare Foundation Model for Clinical Assistance: A Randomized Controlled Trial,” *Nature Medicine* 31 (10) (2025): 3404–3413, <https://doi.org/10.1038/s41591-025-03900-7>.

- ¹⁴ Robert Korom, Sarah Kiptinness, Najib Adan, et al., “AI-Based Clinical Decision Support for Primary Care: A Real-World Study,” *arXiv* (2025), <https://doi.org/10.48550/arXiv.2507.16947>.
- ¹⁵ Eric Topol, *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again* (Basic Books, 2019).
- ¹⁶ Kristine D. Olson, Daniella Meeker, Matt Troup, et al., “Use of Ambient AI Scribes to Reduce Administrative Burden and Professional Burnout,” *JAMA Network Open* 8 (10) (2025): e2534976, <https://doi.org/10.1001/jamanetworkopen.2025.34976>; and Majid Afshar, Mary Ryan Baumann, Felice Resnik, et al., “A Pragmatic Randomized Controlled Trial of Ambient Artificial Intelligence to Improve Health Practitioner Well-Being,” *NEJM AI* 2 (12) (2025): A10a2500945, <https://doi.org/doi:10.1056/A10a2500945>.
- ¹⁷ Nelson Barrera, Maria Solorzano, Yomary Jimenez, et al., “Accuracy of Smartwatches in the Detection of Atrial Fibrillation,” *JACC: Advances* 4 (11.1) (2025): 102133, <https://doi.org/doi:10.1016/j.jacadv.2025.102133>.
- ¹⁸ Teddy Rosenbluth and Maggy Astor, “Empathetic, Available, Cheap: When A.I. Offers What Doctors Don’t,” *The New York Times*, November 17, 2025, <https://www.nytimes.com/2025/11/16/well/ai-chatbot-doctors-health-care-advice.html>; and Brittany Trang, “Patients Are Already Using AI. How Can the Health Care System Help?” *STAT*, November 26, 2025, <https://www.statnews.com/2025/11/26/patients-artificial-intelligence-health-care-system-ai-prognosis>.
- ¹⁹ Rosenbluth and Astor, “Empathetic, Available, Cheap.”
- ²⁰ Alastair Howcroft, Amber Bennett-Weston, Ahmad Khan, et al., “AI Chatbots Versus Human Healthcare Professionals: A Systematic Review and Meta-Analysis of Empathy in Patient Care,” *British Medical Bulletin* 156 (1) (2025), <https://doi.org/10.1093/bmb/ldaf017>.
- ²¹ Eric Topol, “Why Doctors Should Organize,” *The New Yorker*, August 5, 2019, <https://www.newyorker.com/culture/annals-of-inquiry/why-doctors-should-organize>.
- ²² A. Ali Heydari, Ken Gu, Vidya Srinivas, et al., “The Anatomy of a Personal Health Agent,” *arXiv* (2025), <https://arxiv.org/abs/2508.20148>.
- ²³ Eric Topol, “The Emergence of Protein Organ Clocks: Predicting and Tracking Organ-Specific Aging, Health and Disease,” *Ground Truths*, July 13, 2024, <https://erictopol.substack.com/p/the-emergence-of-protein-organ-clocks>.
- ²⁴ Eric Topol, “The Big Shift in Cardiology to Atheroma and Inflammation,” *Ground Truths*, December 3, 2023, <https://erictopol.substack.com/p/the-big-shift-in-cardiology-to-atheroma>.
- ²⁵ Eric Topol, “Predicting and Preventing Alzheimer’s Disease,” *Ground Truths*, May 31, 2025, <https://erictopol.substack.com/p/predicting-and-preventing-alzheimers>; and Eric Topol, “The Breakthrough Blood Test for Alzheimer’s Disease,” *Ground Truths*, April 13, 2025, <https://erictopol.substack.com/p/the-breakthrough-blood-test-for-alzheimers>.
- ²⁶ Eric Topol, “Your Lab Tests: What Is Being Missed by Relying on Population-Based Reference Values,” *Ground Truths*, December 14, 2024, <https://erictopol.substack.com/p/your-lab-tests>.
- ²⁷ Davide Placido, Bo Yuan, Jessica X. Hjaltelin, et al., “A Deep Learning Algorithm to Predict Risk of Pancreatic Cancer from Disease Trajectories,” *Nature Medicine* 29 (5) (2023): 1113–1122, <https://doi.org/10.1038/s41591-023-02332-5>.

- ²⁸ Sadia H. Zapp, “A Breakthrough in Breast Cancer Prevention : FDA Clears First AI Tool to Predict Risk Using a Mammogram,” Breast Cancer Research Foundation, June 2, 2025, <https://www.bcrf.org/blog/clarity-breast-ai-artificial-intelligence-mammogram-approved>.
- ²⁹ Artem Shmatko, Alexander Wolfgang Jung, Kumar Gaurav, et al., “Learning the Natural History of Human Disease with Generative Transformers,” *Nature* 647 (8088) (2025): 248–256, <https://doi.org/10.1038/s41586-025-09529-3>.
- ³⁰ Eric Topol and Kai-Fu Lee, “It Takes a Planet,” *Nature Biotechnology* 37 (8) (2019): 858–861, <https://doi.org/10.1038/s41587-019-0214-z>.
- ³¹ Yu Gu, Jingjing Fu, Xiaodong Liu, et al., “The Illusion of Readiness : Stress Testing Large Frontier Models on Multimodal Medical Benchmarks,” arXiv (2025), <https://arxiv.org/abs/2509.18234>.
- ³² Narges Razavian, Prem Batchu-Green, Vikas Chowdhry, et al., “Catalyzing Health AI by Fixing Payment Systems,” *NEJM AI* 2 (12) (2025): AIpc2500871, <https://doi.org/doi:10.1056/AIpc2500871>.

The Role of AI in Drug Discovery in Africa

Kelly Chibale

The proliferation of artificial intelligence presents Africa's drug discovery landscape with a transformative opportunity to accelerate research and development (R&D) to address the continent's heavy disease burden. The integration of AI in drug discovery processes will help researchers by streamlining research pipelines and improving the design of clinical trials so that historically underserved populations can access cost-effective medication tailored to their needs. While challenges include AI skepticism, scarcity of data, and concerns about job losses, sustained advocacy and capacity-building can help to allay these concerns and facilitate widespread buy-in and adoption of AI tools across the continent. At a time of diminished funding for R&D globally, AI-assisted drug discovery – with its potential to alter fundamentally the pace and cost of R&D – should help Africa move toward a sustainable and innovative pharmaceutical paradigm led by Africans, out of Africa, for the global community.

Home to nearly one-fifth of the world's population, Africa bears a vastly disproportionate burden of infectious disease. This is best exemplified by malaria. The World Health Organization documented 263 million cases of malaria and 600,000 malaria deaths in 2023.¹ Africa accounted for about 95 percent of all malaria cases and mortalities, with African children under the age of five accounting for about 75 percent of all malaria deaths in the region. Other endemic diseases such as tuberculosis, HIV/AIDS, and neglected tropical diseases remain prevalent, and rising rates of noncommunicable diseases such as diabetes and hypertension create a double burden on fragile health systems. This means that Africa faces a unique set of health challenges that both mirror and diverge from those in other parts of the Global South.

The prevention and control of this disease burden in Africa are crucial if the world is to meet Sustainable Development Goal 3 (SDG 3, “to ensure healthy lives and promote well-being for all at all ages”) and especially Target 3.3, which focuses on communicable diseases.² Meeting SDG 3 requires accelerated research and development (R&D) into lifesaving, innovative pharmaceutical interventions that

are cost-effective and can be deployed across the continent at scale, together with urgent improvements to underfunded health infrastructure, fragmented data systems, and issues of access. However, due to a limited capacity to address these challenges locally, Africa has historically been compelled to rely on pharmaceutical interventions primarily from the Global North.

Reducing dependency on international aid and handouts is an increasingly prominent theme in the development discourse across Africa. While drug discovery research efforts have grown locally, they are still in a nascent stage, lacking a critical mass of skills and infrastructure. The challenge has been compounded by the continuous loss of talent from the continent (the “brain drain”) and limited investment in R&D by African governments, along with a gradual decline in contributions from donor countries that has thrown many research projects and clinical trials out of gear. Unlike regions in Asia or Latin America that have made more sustained investments in biotechnology and digital health, Africa’s health innovation ecosystem remains underdeveloped, leading to its underrepresentation in global pharmaceutical R&D.

We are at a critical juncture as well as unique moment of opportunity in Africa’s history. With traditional channels of aid and support disrupted, Africa must rely on its own capabilities and build upon them to deliver cutting-edge, innovative medicines to the people who need it most.

Enter artificial intelligence. Even in its infancy, AI has had a transformative impact across virtually every sector of society. Consider its impact – and the recognition of it – in scientific research. In 2024, John Hopfield and Geoffrey Hinton were awarded the Nobel Prize in Physics for their work on artificial neural networks in machine learning (ML).³ That same year, Demis Hassabis and John Jumper were awarded the Nobel Prize in Chemistry for their contributions in developing an AI tool, Google DeepMind’s AlphaFold, that can predict the complex three-dimensional structures of proteins from the primary amino acid sequence. Apart from AlphaFold, David Baker also shared the Nobel Prize in Chemistry for his contributions to novel protein design.⁴

AI has the potential to accelerate health innovation and tailor drug discovery to Africa’s disease burden. This is especially exciting in contexts like malaria or sickle cell disease, for which local prevalence is high and global pharmaceutical investment is limited. While its applications in drug discovery are incremental, AI is already catalyzing efforts to develop a malaria vaccine, combat drug-resistant bacteria, and mine local genomic and health data for new therapeutic insights.⁵ These advances are possible because AI has the ability to identify patterns in data, process much larger datasets than humans, and automate complex tasks. While drug discovery is traditionally a time-consuming and expensive process, AI can potentially accelerate every step of the process, saving time and vast resources while providing novel insights.

Africa owes its limited capacity for drug discovery and development to several historical factors. One is the reliance on handouts and assistance mostly from the Global North. The continent is now enduring major disruptions to R&D in the wake of diminished donor funding. The problem is compounded by the historical general lack of investment in infectious disease R&D by African governments, the African private sector, as well as most innovative pharmaceutical companies due to low anticipated commercial returns.

Inequity in the global health landscape also puts Africa at a disadvantage, as the COVID-19 pandemic made clear. During the global rollout of vaccines, Africa received a minuscule allocation while countries in the Global North hoarded multiple doses for their citizens. By October 2022, nearly three years after the pandemic began, Africa had received only 540 million doses, or 6 percent of the 9 billion produced worldwide, with less than 10 percent of its population fully vaccinated by that time.⁶

This inequity also manifests itself in the paucity of clinical trials in Africa, which stood at 2.2 percent of all trials conducted worldwide between 2012 and 2023, despite the continent bearing 20 percent of the world's disease burden and greater genetic diversity than the rest of the world combined.⁷ The absence of data on how drug dosages affect African genetics has led to serious consequences for the continent and its people. For example, the HIV/AIDS drug Efavirenz was launched as a first-line treatment in Zimbabwe in 2015 after being used successfully in the United States and Europe. But Zimbabwean HIV/AIDS patients experienced serious side effects that are associated in part with genetic mutations in the gene encoding for the enzyme responsible for the metabolism of Efavirenz. These mutations lead to a slow metabolizer phenotype, implying that Zimbabwean patients were effectively being overdosed, which accounts for the side effect profile in these patients. Dosing recommendations were based on the fact that such mutations are rare in Europeans and Americans (as low as 3 percent) but found in 20 percent of Africans.⁸ Addressing such gaps will require including African data in drug discovery and development.⁹ It is critical, therefore, for Africa to build its own drug discovery research capabilities to develop solutions for its most pressing disease challenges.

This is easier said than done. Financial constraints in low- and middle-income countries (LMICs) stand in the way of adequate infrastructure, including laboratories equipped with advanced technologies. Funding shortfalls result in a shortage of trained researchers and health care professionals. Drug discovery is also an inherently lengthy and costly process fraught with risk of failure. The journey of a drug from the laboratory to the shelf often takes a decade, if not more, with appropriate sustained investment. In Africa's case, the writing is on the wall. The challenges it faces are immense, but twenty-first-century tools like AI provide the continent with a unique opportunity to chart its own innovative pathways to drug discovery – by Africans, out of Africa, for the global community.

The Holistic Drug Discovery and Development (H3D) Center, Africa's first and leading world-class integrated drug discovery and development center, has been spearheading the continent's push for self-reliance in drug discovery.¹⁰ I founded H3D at the University of Cape Town in 2010, and have directed it and the associated H3D Foundation since 2019. H3D works with a range of partners and collaborators to reduce barriers for African scientists, break new ground in the discipline of drug discovery, and expand the drug discovery ecosystem and community in Africa. H3D collaborated with the African Academy of Sciences, the Bill & Melinda Gates Foundation through its Grand Challenges initiative, and the Medicines for Malaria Venture to identify the best emerging African scientists and build a cohort of future leaders within Africa, thus strengthening the drug discovery discipline and mitigating the brain drain from the continent. This led to the initiation of the Grand Challenges Africa Drug Discovery Sentinel program in 2018, with H3D providing direct support to all the grantees.

In November 2023, the H3D Foundation, with \$7.2 million in funding from the Gates Foundation and LifeArc, established the Grand Challenges African Drug Discovery Accelerator (GC ADDA), a network of drug discovery researchers working in Africa on unmet medical needs faced by the African patient.¹¹ By 2024, the network had already linked eight African countries and twenty-one research institutions. The GC ADDA seeks a future in which African patients have access to cost-effective, efficacious disease treatments tailored to their health needs. With four initial flagship projects, GC ADDA is working to create regional centers of excellence within the continent to grow and sustain self-reliant research in drug discovery.

The H3D Center's focus on innovation to develop platform technologies positions it as an early adopter of AI in drug discovery research. In 2021, H3D entered a partnership with the Ersilia Open Source Initiative (EOSI) with a mission to incorporate AI and ML tools into the H3D drug discovery pipeline.¹² This was an ideal partnership: H3D at that time had no in-house AI/ML expertise, and EOSI specialized in developing AI/ML tools for drug discovery in LMICs. In 2023, H3D received a major boost for AI-enabled drug discovery when I was awarded a Schmidt Sciences AI2050 Senior Fellowship through the H3D Foundation. The AI2050 work involves linking open-source AI to existing pharmacometrics (mathematical) modeling tools used in human dose prediction for clinical trials and boosting the development of ready-to-use AI models for malaria and tuberculosis drug discovery in Africa.¹³

Traditional methods of drug discovery are dependent upon an approach of examining large numbers of potential drug candidates and sorting through them to identify "lead" compounds. This is a cumbersome, expensive process, hamstrung by the limited availability of these compounds and the challenge of predicting how they will interact with the human body. Because AI algorithms can be applied across thematic domains and trained on a broad spectrum of data-

sets, they can be employed across different stages of the drug discovery pipeline to make the process much more time- and cost-effective. An example of this would be the chemistry-related or protein-related models that have been trained on network architectures inspired by large language models such as ChatGPT.

Training AI models on large datasets enables them to identify patterns in diseases and predict the potential composition of drugs that can treat them. Resources like AlphaFold, which can predict the 3D structure of proteins, can help accelerate this process. AI's ability to leverage statistical algorithms to anticipate how drug candidates will bind to specific targets and what properties they will have can also reduce the need for – and the costs associated with – physical testing of those compounds. Chemists can thus examine a much wider set of prospective compounds much more efficiently. AI-assisted approaches have also been known to rank promising drug candidates more accurately for further assessment compared with traditional methods.¹⁴

I had the privilege of being invited by the Wellcome Trust to join a Scientific Advisory Committee of about fifteen leaders in drug discovery across industries and geographies to offer input on and guide research led by the Boston Consulting Group for a new landscaping report on the impact of AI in drug discovery.¹⁵ This report looked to understand at which points in drug discovery AI was being deployed, identify new areas to expand its use, and assess the extent of adoption of these tools across the drug discovery ecosystem. There was a strong focus on understanding where AI can be expanded across infectious disease drug development and how to increase adoption in LMIC settings.

A leading use case of AI for small molecule research is in the development of quantitative structure–activity relationship (QSAR) and quantitative structure–property relationship (QSPR) prediction models to identify promising compounds prior to synthesis and increase the efficiency of the experimental pipeline. A prominent example of this approach is the end-to-end, fully automated implementation of an AI/ML tool for QSAR/QSPR modeling called ZairaChem. With no prior AI/ML capabilities, the H3D Center implemented ZairaChem, a first-of-its-kind deployment at scale of AI/ML tools in a research center operating in a low-resource setting.¹⁶ With a light requirement for computational resources, the tool assisted H3D in developing a virtual screening cascade for malaria and tuberculosis drug discovery. Trained on the H3D Center's in-house compound library, the virtual screening cascade consists of a set of classification models that predict whether a given compound is likely to be “active” or “inactive” across various measured biological properties. Far from replacing lab-based scientists, expert insights from experimentalists are critical in developing relevant and accurate models. The AI models undergo retraining every six months to ensure they are updated with the most recent in-house data and remain aligned with ongoing research programs. Models can also be developed on request for specific drug discovery programs or for additional datasets of inter-

est. The periodic maintenance of AI models, including engaging with lab-based scientists during model development and regular upskilling of scientists in the use and interpretation of model outputs, is critical to maximizing the potential of AI in drug discovery research.¹⁷ ZairaChem is now being applied to Gram-negative and Gram-positive pathogens responsible for antibiotic-resistant microbial diseases that are of global interest, given the threat of antimicrobial resistance.

More advanced applications include the use of generative AI techniques in the *de novo* design of small molecules. Considering the vastness of small-molecule chemical space, AI facilitates the exploration of this chemical space to generate safe and efficacious molecules. Generative AI methods produce vast libraries of virtual compounds that need to be logically filtered to a more manageable subset that balances chemical novelty with relevance to the project; the H3D Center has begun exploring these methods in programs across its different disease areas. Multimodal AI models link different data sources, such as images with assay data, to improve the accuracy of model predictions or answer new research questions. For example, image classification models can be trained on microscopy images to identify cell phenotypes associated with inhibiting specific proteins, an approach the H3D Center is exploring to understand potential malaria treatments. Thus, AI has enhanced researchers' ability to design drugs more efficiently and accurately and better understand disease biology.

Further AI applications include pharmacogenomics research, in which sophisticated AI tools are utilized to customize drug dosages and predict treatment outcomes based on genetic markers of specific diseases. This translates to a more personalized treatment of disease, which is vital for the African context. Despite being the most genetically diverse continent, Africa's severe paucity of clinical trials, as discussed earlier, has led to suboptimal treatment outcomes and raised questions about dose optimization for its population. AI-assisted data analysis of the clinical data landscape in Africa can improve clinical trial design and help increase the efficacy of clinical trials.¹⁸ In these ways, AI can help to tailor drugs for the genetically diverse African population.¹⁹

At H3D, our initial contribution has been to apply a transfer-learning approach to mitigate the low-data paradigm and to identify relevant pharmacogenes for a given drug molecule. To do this, we integrated clinical and preclinical drug and gene data available in the public domain (that is, from prior experiments conducted globally). The Project Africa GRADIENT (Genomic Research Approach for Diversity and Optimizing Therapeutics) initiative explores the influence of the continent's genetic variability on existing therapies, especially for malaria and tuberculosis.²⁰ H3D's project within this initiative looked at the potential of AI/ML in pharmacogenomics to predict dosing adjustments for some drugs from a list of over thirty malaria and tuberculosis drugs prescribed in Africa. The predicted metabolism-related genes could then refine pharmacometrics models for human dose predictions to

optimize treatment for an individual's genetic makeup.²¹ Genetic data collection conducted across the continent should greatly advance pharmacogenomics research in Africa enormously.

Despite the immense potential and noteworthy accomplishments of AI-assisted drug discovery research, there are significant challenges that stand in the way of its acceptance, adoption, and uptake. And within this context, there are the Africa-specific and major challenges of internet connectivity, low AI literacy, and limited digitalization. Further, the dominant discourse around AI is polarized, with sharp divides regarding the ethics of how AI models are trained, how they interact with intellectual property (IP) rights, and what tasks AI will “take over,” possibly costing people their livelihoods.

The hype surrounding AI and the debates it spawns leave us with a paradigm in which many intended end users, including scientists, are skeptical of AI, if not hostile to it. End users are often not computationally trained, and their uptake of AI is hindered by a poor understanding of how to leverage insights from AI, as well as a lack of trust in predictions typically produced by “black box” models. It is reasonable to surmise that skepticism around AI is more pronounced and/or exacerbated in Africa due to the continent's history and limited access to the basic infrastructural ingredients that underpin AI.

An added challenge is that interaction with AI is fundamentally different from how society is accustomed to interacting with standard computer algorithms. Traditionally, computers have been used as “calculators” that follow a predefined formula to arrive at an output. However, AI output is intrinsically more probabilistic and leverages past learnings and context to make predictions. These are not – and will not always be – correct for every given input, but they still provide value at a larger scale over many molecules. Iterative testing can help bridge this lack of trust.

The biggest fear surrounding the proliferation of AI tools is that they will replace jobs. Though the experience in drug discovery so far has clearly shown that AI tools can neither substitute entirely for traditional research methodologies nor replace researchers' expertise, AI tools offer predictions and insights subject to the data they are provided with. Making sense of these insights in the context of the research requires human intervention. Skeptics have also suggested that scientists who do not embrace AI could lose their jobs. But we have already seen AI facilitate job creation, as in the context of digital health care. For Indian hiring companies, health informatics and medical AI are rapidly becoming two of the most promising career paths, evidenced by the number of informatics jobs created by hospitals, startups, government departments, and individuals who studied in technology fields.²² The *complementary* design-make-test-analyze approach used for small molecule drugs – which takes advantage of AI's cost-effective, time-saving properties through the

rapid analysis of vast datasets or chemical spaces – alongside the skills and knowledge of human researchers is the way forward to optimizing drug discovery research.

Another major challenge concerns data. Researchers working with AI tools rely heavily on the availability of data to train their models and derive insights. While pharmaceutical datasets are vast, they are often severely access-limited. Datasets that are publicly available (such as clinical trial data) may be of poor quality, as they often are more biased due to being collected in a more-restricted context, and are smaller and heterogeneous, since they come from multiple sources. AI tools trained on such noisy data are susceptible to inaccurate predictions; this limits the applications of AI and the development of novel tools that could accelerate research. It could also lead to the development of drugs with limited efficacy in certain populations, as many African countries lack comprehensive biobanks, standardized health records, and data-sharing infrastructure. Existing datasets often reflect non-African populations, leading to models that may not generalize well in African contexts. There are also challenges around data governance, including ethical concerns, patient consent, and the risk of data exploitation. Addressing these issues requires investment in local data infrastructure, research capacity, and policies that promote equitable data ownership and sharing.

While the latest literature reports many new AI-driven drug discovery research tools, there is often little funding incentive to make these tools easily accessible, open-source, or regularly maintained through periodic retraining. As a result, the scientific literature on the development of AI tools for drug discovery grows rapidly with little emphasis on adoption of these tools postpublication.

EOSI is actively taking on this challenge with its Ersilia Model Hub. By disseminating AI models and published findings to researchers in the Global South, they enable the improvement of drug discovery efforts in low-resource regions. Most of the Hub's computational models are open-source and publicly available, contributed largely by data scientists and software engineers, and most users belong to drug discovery institutions across sub-Saharan Africa. Because they allow researchers to computationally simulate the interaction between a drug candidate and a specific disease pathogen, these models are cost-effective. H3D is EOSI's biggest partner in Africa.

As with all new technologies, the question of ethics is crucial. We must strike a balance between patient rights and the use of technology when we incorporate AI tools into drug discovery research. Sensitive data collected from patients for clinical trials must be treated with the utmost care and anonymity to ensure their privacy is not compromised. Informed consent is also essential to mitigate the continent's historical distrust of clinical trials. Regular audits of AI algorithms and their updates with the latest and most diverse datasets can address the challenge of bias in data that, as discussed, can reduce the efficacy of drugs produced for underserved populations. This is even more important in Africa. IP protection of drug discovery datasets is equally important. We need to find ways to incentivize

industry and African partners to share their datasets, if not publicly, then at least to train AI models for the benefit of the global community. It would be inequitable for Global North companies to retain their data confidentiality while African institutes are expected to relax their data confidentiality for the sake of AI progress.

These actions require collaboration between researchers, health care providers, data scientists, and software developers. Such practices help to build confidence in clinical trials and foster greater participation, allowing AI to enter the public health domain in an environment of positivity and trust. After all, new technologies are meant to enhance equity in public health rather than perpetuate existing inequalities and exacerbate social injustices. Hence, without addressing these structural gaps, AI risks reproducing existing inequities rather than resolving them. Alternatively, if developed in a locally relevant and inclusive way, AI could help reposition Africa from a passive recipient of health innovations to an active contributor to global biomedical research.

We are at a pivotal stage in the history of drug discovery and public health. The transition from traditional research methods to AI-assisted research will provide scientists and researchers with tools that catalyze processes and achieve better results. With the assistance of AI, the coming decades could see the Global South, especially Africa, make great progress to level the global playing field of pharmaceutical innovation and give the African people the cost-effective, efficacious treatments that they need. And while this essay has not placed much emphasis on generative chemistry and associated synthesis planning or on lab automation, all of which AI can facilitate, these can have critical importance in Africa, where the medicinal chemistry capacity is smaller than in the Global North. Furthermore, given the long history in Africa of exploring products from various natural sources, including African traditional medicines, for drug discovery, there is a tremendous opportunity to integrate AI into African natural products research. For example, AI can help to predict the activity of natural products like those derived from plants widely used in African traditional medicines, elucidate the composition of lab extracts, and help “simplify” natural product structures so they are more synthetically accessible.

Extensive training and consultations are the first steps toward growing literacy around AI tools and enabling the subsequent adoption of AI technologies. Ironically, it is this “human interaction” that minimizes the perceived barriers to the adoption of AI-based insights. In this regard, capacity-building activities such as workshops are crucial as they can aid the emerging networks of African scientists. This has already been demonstrated by workshops run jointly by the H3D Foundation and EOSI through my Schmidt Sciences AI2050 Senior Fellowship.

As AI becomes more prevalent in modern society, the importance of large-scale and high-quality data is becoming increasingly apparent. AI models need to be

trained on diverse and representative datasets to ensure biogeographical and/or ethnic representativeness. Truly democratized access to the benefits of AI will first require an increase in democratized access to data. These models must also be periodically reviewed and updated with newer data to ensure that their intervention follows the evolving realities of the infectious disease scenario in Africa.

Sensitive patient data such as medical histories and genetic information need to be protected to build wider trust in AI-assisted drug discovery processes, especially among governments, the media, and the larger public. Robust regulatory and ethical frameworks must consider how AI influences drug discovery processes to ensure that medicines developed with AI support continue to be held to the same rigorous standards as those that have been developed traditionally. “Federated learning” or “swarm learning,” in which multiple institutions (such as hospitals) use their data to train an AI model jointly in a privacy-protected way, should be encouraged.

Historically, Africa has often been marginalized in global biomedical research, with its role largely limited to sites for occasional clinical trials, rather than as a hub for innovation. However, recent shifts – such as the growth of African genomics initiatives and the continent’s increasing digital connectivity – signal a potential turning point. Looking ahead, the integration of AI into drug discovery offers a chance not only to leapfrog infrastructural limitations but also to redefine Africa’s role in shaping global health solutions. If aligned with long-term investments in education, data sovereignty, and equitable research partnerships, the next five to ten years could see Africa moving from being underrepresented in the medical research pipeline to playing a leadership role in developing treatments for both local and global health challenges.

Around the world, AI has forced professionals to fundamentally rethink the way they approach their work. In science, AI has shown – and realized – great potential as a tool that augments the intellect and rigor of the scientist; a companion that catalyzes processes and allows humans to devote more time and thought to the problems they are trying to solve. Its relative cost-effectiveness democratizes scientific research and opens the door for researchers in the Global South, especially in Africa, to participate in the cutting-edge of drug discovery and achieve positive change for the people that need it the most.

The story of AI in African drug discovery is not only about present capabilities but also about reclaiming agency over future biomedical innovation. AI can be of interest in Africa for various applications, including better characterization of pathogen biology (in particular, to discover new essential/vulnerable proteins in poorly studied organisms), identification of drug repurposing opportunities, prediction of pathogen drug resistance, and elucidation of active natural products from medicinal plants.

AUTHOR'S NOTE

I am grateful to James Manyika for the invitation and opportunity to contribute this essay and for his guidance during its preparation. I have benefited from core ideas discussed with Jason Hlozek (H3D/UCT) and John Woodland (H3D/UCT), who, along with Miquel Duran-Frigola (Ersilia), provided feedback on an early draft. I would like to thank Dineo Moerane and her colleagues at Global Health Strategies for their help with converting the core ideas into a presentation format.

ABOUT THE AUTHOR

Kelly Chibale is Professor of Organic Chemistry at the University of Cape Town (UCT), where he holds the Neville Isdell Chair in African-Centric Drug Discovery & Development. He is also the Founder and Director of the UCT Holistic Drug Discovery and Development (H3D) Center, the Founder and CEO of the H3D Foundation, a Schmidt Sciences AI2050 Senior Fellow, a Full Member of the UCT Institute of Infectious Disease & Molecular Medicine, and the Founding Director of the South African Medical Research Council Drug Discovery and Development Research Unit at UCT. He serves as Editor-in-Chief of the American Chemical Society's *ACS Medicinal Chemistry Letters*.

ENDNOTES

- ¹ World Health Organization, *World Malaria Report 2024: Addressing Inequity in the Global Malaria Response* (World Health Organization, 2024).
- ² World Health Organization, "SDG Target 3.3 Communicable Diseases," https://www.who.int/data/gho/data/themes/topics/sdg-target-3_3-communicable-diseases (accessed July 28, 2025).
- ³ Nobel Foundation, "Nobel Prize in Physics 2024," <https://www.nobelprize.org/prizes/physics/2024/summary> (accessed July 28, 2025).
- ⁴ Nobel Foundation, "Nobel Prize in Chemistry 2024," <https://www.nobelprize.org/prizes/chemistry/2024/summary> (accessed July 28, 2025).
- ⁵ Google DeepMind, "Stopping Malaria in Its Tracks," October 13, 2022, <https://deepmind.google/discover/blog/stopping-malaria-in-its-tracks> (accessed July 28, 2025); and Felix Wong, Erica J. Zheng, Jacqueline A. Valeri, et al., "Discovery of a Structural Class of Antibiotics with Explainable Deep Learning," *Nature* 626 (7997) (2024): 177–185.
- ⁶ Michel Sidibé, "Vaccine Inequity: Ensuring Africa Is Not Left Out," Brookings Institution, January 24, 2022, <https://www.brookings.edu/articles/vaccine-inequity-ensuring-africa-is-not-left-out> (accessed July 28, 2025).
- ⁷ Avik Pal, "Counting Africa in for Clinical Trials," *Forbes*, September 24, 2024, <https://www.forbes.com/councils/forbesbusinesscouncil/2024/09/24/counting-africa-in-for-clinical-trials>.
- ⁸ Andrew Park, Daniel Steel, and Elicia Maine, "Evidence-Based Medicine and Mechanistic Evidence: The Case of the Failed Rollout of Efavirenz in Zimbabwe," *The Journal of Medicine and Philosophy* 48 (4) (2023): 348–358.

- ⁹ Clinton G. L. Veale, Adrienne L. Edkins, Susan Winks, Mathew Njoroge, and Kelly Chibale, “Including African Data in Drug Discovery and Development,” *Nature Reviews Drug Discovery* 22 (7) (2023): 521–522.
- ¹⁰ The Holistic Drug Discovery and Development (H3D) Center, University of Cape Town, <https://h3d.uct.ac.za>; and Ari Daniel, “This Lab That’s Determined to Discover New Drugs Isn’t Where You Might Expect,” NPR, March 22, 2026, <https://www.npr.org/2026/03/22/g-s1-114038/new-drugs-health-medicine-africa-tuberculosis-measles>.
- ¹¹ LifeArc, “LifeArc Makes Multi-Million Pound Investment to Support Drug Discovery in Sub-Saharan Africa,” GlobeNewswire, January 31, 2024, <https://www.globenewswire.com/news-release/2024/01/31/2820723/0/en/LifeArc-makes-multi-million-pound-investment-to-support-drug-discovery-in-sub-Saharan-Africa.html> (accessed July 28, 2025).
- ¹² Ersilia Open Source Initiative, “Ersilia,” <https://www.ersilia.io> (accessed July 28, 2025).
- ¹³ University of Cape Town, “Second Cohort of AI2050 Senior Fellows Named,” October 20, 2023, <https://www.news.uct.ac.za/article/-2023-10-20-second-cohort-of-ai2050-senior-fellows-named>.
- ¹⁴ Matthew Chun, “How Artificial Intelligence Is Revolutionizing Drug Discovery,” Petrie-Flom Center of Health Law Policy, Biotechnology, and Bioethics at Harvard Law School, March 20, 2023, <https://petrieflom.law.harvard.edu/2023/03/20/how-artificial-intelligence-is-revolutionizing-drug-discovery>.
- ¹⁵ Wellcome Trust, *Unlocking the Potential of AI in Drug Discovery: Current Status, Barriers and Future Opportunities* (Wellcome Trust, 2023).
- ¹⁶ Gemma Turon, Jason Hlozek, John G. Woodland, et al., “First Fully-Automated AI/ML Virtual Screening Cascade Implemented at a Drug Discovery Centre in Africa,” *Nature Communications* 14 (1) (2023): 5736.
- ¹⁷ Jason Hlozek, Kelly Chibale, and John G. Woodland, “Ongoing Implementation and Prospective Validation of Artificial Intelligence/Machine Learning Tools at an African Drug Discovery Center,” *ACS Medicinal Chemistry Letters* 15 (7) (2024): 989–993.
- ¹⁸ Ruishan Liu, Shemra Rizzo, Samuel Whipple, et al., “Evaluating Eligibility Criteria of Oncology Trials Using Real-World Data and AI,” *Nature* 592 (7855) (2021): 629–633.
- ¹⁹ Gemma Turon, Mathew Njoroge, Mwila Mulubwa, et al., “AI Can Help to Tailor Drugs for Africa—but Africans Should Lead the Way,” *Nature* 628 (8007) (2024): 265–267.
- ²⁰ Carene Anne Alene Ndong Sima, Houcemeddine Othman, Marlo Möller, and The Project Africa GRADIENT Consortium, “Advancing Pharmacogenetics Research in Africa: The ‘Project Africa Gradient’ Initiative,” *Drug Discovery Today* 29 (5) (2024), <https://doi.org/10.1016/j.drudis.2024.103939>.
- ²¹ Mwila Mulubwa and Kelly Chibale, “Model-Informed Approaches to Optimizing Therapeutics in the African Patient Populations,” *Clinical Pharmacology and Therapeutics* 116 (6) (2024): 1391–1394; and Gemma Turon, Mwila Mulubwa, Anna Montaner, et al., “AI Coupled to Pharmacometric Modelling to Tailor Malaria and Tuberculosis Treatment in Africa,” *Nature Communications* (2025), <https://doi.org/10.1038/s41467-025-64304-2>.
- ²² Nandita Vijayasimha, “Health Informatics and Medical AI Are Making Waves in Job Market,” Pharmabiz, April 28, 2025, <https://pharmabiz.com/NewsDetails.aspx?aid=177570&sid=1> (accessed July 28, 2025).

Physics Is Different: Context, Culture & Craft in Effective AI for Physics

Tess Smidt

Physics is a sprawling and ambitious enterprise unified by the goal of explaining the natural world, yet it is extraordinarily diverse in how that goal is pursued. The opportunities to harness artificial intelligence to aid in our pursuits to understand the universe are vast. But the heterogeneity of physics in practice is a vivid demonstration of how scientific inquiry varies across scale, infrastructure, and standards of evidence, and how machine learning integrates unevenly as a result. While current applications of AI often focus on accelerating existing workflows, the deeper promise lies in developing robust methods that plug into physicists' established workflows and toolkits; lower barriers in the continuous cycle between measurement, simulation, analysis, and theory; and create fundamentally new ways of probing, modeling, and controlling physical systems.

Physics is big. Really big. It is both the science of first principles and the practice of probing every corner of the natural world, from subatomic particles to the cosmos. The word itself suggests unity, but in practice the field is anything but monolithic. Particle physicists build colliders the size of cities to glimpse fleeting quarks and gluons. Condensed matter physicists coax exotic behaviors out of crystals grown milligram by milligram. Astrophysicists study a universe that can never be rerun. Plasma physicists wrestle with the controlled chaos of hot, charged matter. What unites these efforts is an ambition to explain how the world works, and a shared toolbox of habits – reductionism, emergence, approximation, symmetry, unlearning – that make the complexity of nature tractable.

How artificial intelligence enters these pursuits is bespoke and uneven. Artificial intelligence techniques already accelerate simulations, parse noisy detector streams, and help steer experiments in real time. In some subfields, they are embedded deeply into workflows; in others, their role remains speculative. There is a popular temptation to treat AI as a universal solvent, an inevitable engine of automated discovery. But physics offers a cautionary tale. The field's extraordinary diversity of data, infrastructure, and culture means there is no single interface for machine

learning to plug into. The tools that succeed will amplify human judgment, stay grounded in physical context, and enrich the shared work of interpretation, helping us frame new questions and broaden what counts as inquiry.

This essay is an attempt to map that terrain. I begin by sketching the varied enterprise of physics itself, highlighting the activities that define it – experiment design, measurement, simulation, analysis, theory, and modeling – and the infrastructures that sustain those activities. I then turn to concrete vignettes, from the fragile ripples of gravitational waves to the torrents of data at the Large Hadron Collider (LHC) to the patchwork of experiments and simulations in atomistic science, to show how machine learning is already entangled with physics in practice. From there, I consider the toolboxes of physics and of machine learning and the resonances that have begun to emerge between them. Finally, I explore the challenges and frontiers ahead: what it will take for AI not merely to speed up existing pipelines, but to help us articulate new abstractions, reframe problems, and expand what it means to understand.

The message I hope to leave you with is this: physics will not be “solved” by AI. But AI can become an essential part of the physicist’s toolkit, lowering barriers between experiment design, measurement, simulation, and modeling, and giving us new knobs to probe and play with the universe.

Physics has a reputation for elegance, compact equations like $E = mc^2$ or $F = ma$ that seem to distill reality into symbols. But behind those icons lies an extraordinary range of practices. Doing physics means designing experiments and instruments, making measurements, running simulations, analyzing data, and modeling phenomena. No single physicist traverses all of these steps; the field is sustained by collaborations in which each specialist contributes deep expertise to a narrow part of the cycle, with just enough overlap to communicate across boundaries.

These activities are worth outlining not only to show the variety of what physicists do but also to set the stage for how artificial intelligence might help at each step.

Design. Often, physicists must first create the very systems they hope to study. This can mean assembling lattices of ultracold atoms on an optical table or growing crystals grain by grain in a lab furnace. It also includes design of the experimental tools needed to capture signals, such as fabricating detectors that hold thousands of tons of liquid argon to entice elusive neutrino interactions or kilometers of arrays of antennas looking at the spectra of the cosmos.¹ Experiment design pushes the limits of engineering to expand what questions can even be asked; what we can build shapes the scope of inquiry long before any measurement is made.

Measurement. Measurement captures signals from the world, whether naturally occurring or carefully contrived. At one extreme are kilometer-scale interferometers like the Laser Interferometer Gravitational-Wave Observatory (LIGO) that

detect spacetime ripples a thousand times smaller than a proton's width.² At the other extreme are tabletop spectrometers that fit on an optical bench and reveal atomic transitions through the absorption or emission of light. To a physicist, there is no hard line between system and instrument: every observation is an interaction between physical actors. Even the simplest measurements are mediated by probes (made of photons, electrons, neutrons, and so on) that shape the result.

Simulation. When direct measurement is impossible, physicists turn to simulation: structured guesswork that explores what might happen under given laws. Early Monte Carlo methods, invented at Los Alamos, were designed to track neutrons through fission reactions.³ Today, simulations extend across scales: particle physics detectors modeled in exquisite detail; universes evolved under different cosmological parameters; atomic trajectories stepping through femtoseconds of molecular motion. Simulations are not “truth machines” but approximations, balancing fidelity against computational cost. The assumptions built into simulations – what physics to include, what variables to simplify – matter as much as the simulation's outputs.

Analysis. Raw signals or simulated traces rarely speak for themselves. Analysis transforms them into evidence through statistics, signal processing, calibration, and inference. Consider the Large Hadron Collider: thirty million proton-proton collisions per second, most discarded within microseconds by custom electronics and fast classifiers.⁴ Or cosmological surveys, in which galaxies must be disentangled from atmospheric noise, detector artifacts, and foreground contamination. Analysis is where judgment enters most visibly, deciding which patterns are significant, which uncertainties dominate, and which anomalies to trust.

Theory and modeling. Physics is perhaps best known as a science of first principles. Its iconic laws – Maxwell's equations, quantum mechanics, general relativity, the Standard Model – suggest a discipline governed by elegant, universal rules. These are what physicists call theories: compact statements about how the world should behave, often formulated with remarkable mathematical economy.

Yet those pristine equations are rarely sufficient on their own. They describe everything, and therefore almost nothing in particular. Turning theory into a working tool requires modeling, constructing simplified or approximate systems that make the consequences of theory concrete. Modeling translates universal law into a form that can be calculated, compared, and understood. The Ising model distills the essence of phase transitions; Ginzburg-Landau theory captures superconductivity through a coarse-grained order parameter; simplified cosmologies let us test how inflation or dark matter might leave observable traces. A good model makes the overwhelming interpretable, showing which interactions or symmetries truly matter.

Reductionism and emergence guide this process from opposite ends. Modeling often begins by reducing complex phenomena into simpler parts, yet the rich-

est insights appear when new collective behaviors emerge that the fundamental equations never single out. Symmetry sits between them. It describes invariance, the aspects of a system that remain unchanged under transformation; and when symmetry is broken, it reveals how new structure or order can arise. Principles like these help physicists reason about systems through constraints and conservation laws rather than by tracking every microscopic detail.

Approximation is a constant companion for understanding a complex world. Almost no real-world problem can be solved exactly. Physicists linearize equations, expand functions in series, or integrate out degrees of freedom to make problems tractable. In many cases, knowing which details to ignore is as important as knowing which to keep. Yesterday's laws often survive as limiting cases of broader ones: Newtonian mechanics within relativity, classical electrodynamics within quantum field theory.

Theory provides the general rules; modeling provides the playground where those rules become intelligible. Machine learning faces a similar challenge. Neural networks can, in principle, represent extraordinary complexity, but to make sense of what they learn requires added structure and interpretation. This should not be surprising; even in physics, compact laws alone do not yield intuition. What turns theory into insight is the scaffolding that manages complexity and invites play: the approximations, symmetries, and models that let us explore how systems behave when their rules are bent or simplified.

A good scientific tool, in this sense, is not one that simply confirms what we expect but one that has the authority to surprise us. The same must be true for new tools of artificial intelligence. To contribute meaningfully, they cannot bypass the habits that make understanding possible; they must work within them or help us find more powerful paradigms. The most promising uses of AI in physics will be to build tools that amplify and reorganize the familiar cycles of design, measurement, simulation, analysis, theory, and modeling.

Interdependence and infrastructure. Though described separately, these activities (design, measurement, simulation, analysis, theory, and modeling) are tightly intertwined. Simulations depend on models; analyses depend on simulations; measurements depend on design; theories are revised by all. And none of this happens in a vacuum. Infrastructure – whether a shared beamline, a national supercomputer, or a tabletop laser – sets both the feasibility of questions and the barriers to asking them. In particle physics, data volumes are so immense that experiments resemble institutions more than instruments. In condensed matter, the bottleneck may be access to a single synchrotron beamline for a few hours of precious time.

Gravitational waves are ripples in spacetime produced when massive objects accelerate asymmetrically, for example, black holes or neutron stars orbiting each other in an ellipse. Einstein predicted gravitational waves in 1916;

they were first observed a century later, in 2015, by LIGO.⁵ Detecting them requires astonishing precision, measuring changes in distance a thousand times smaller than the width of a proton.

Each of the two LIGO sites consists of a giant four-kilometer-long interferometer that splits a laser beam into two perpendicular arms. A passing gravitational wave stretches one arm while squeezing the other, altering the distance the light travels. For a measurement to count, both sites must be able to measure the same ripple. The challenge lies in the faintness of the signal and the overwhelming background of noise. The detectors are exquisitely sensitive to any vibration, whether it be earthquakes, air conditioners, thermal drift, lightning strikes, even cosmic rays.⁶ Keeping the mirrors aligned and the instrument stable is key for a meaningful measurement.

Experimental time is too valuable to risk on untested methods, so most are proven offline to earn a role in live operations. And because measuring something unprecedented demands precision and interpretability, AI might seem an unlikely tool. But when carefully targeted, machine learning has already proven useful.

For instance, machine learning has been used to improve mirror control, suppressing vibrations and technical noise, and to stabilize the light injection system, particularly the delicate quantum “squeezing” process that reduces photon noise at high frequencies.⁷ These approaches show how AI can enhance critical control systems, extending the detectors’ reach.

Machine learning is also now part of LIGO’s real-time detection pipeline. For example, the Aframe system uses deep neural networks trained on millions of simulated waveforms mixed with real detector noise to identify gravitational-wave signals within seconds of their arrival.⁸ Its performance rivals traditional matched-filter searches while running far faster and with lower computational cost. The companion system AMPLFI applies similar methods to estimate the properties of each event, such as the masses and location of merging black holes, almost immediately.⁹ Together, these AI-based pipelines now issue public gravitational-wave alerts in real time, as for the recent event S250830m, very likely a binary black hole merger.¹⁰

A team of researchers from the Max Planck Institute for the Science of Light and the LIGO Laboratory at Caltech used AI not to run detectors but to reimagine them.¹¹ By searching the immense design space of possible interferometer layouts with a “universal interferometer” model and large-scale physics-based simulations, they discovered dozens of novel topologies that outperform current next-generation designs in sensitivity. While these ideas face major hurdles of cost, engineering, and implementation, they demonstrate how AI can propose new directions for the future of gravitational-wave detection.

A common thread of all these applications of machine learning for gravitational-wave detection is that it is bespoke. While elements might step from other applica-

tions, the end-to-end algorithm (the input data, architecture, loss functions, and training scheme) needed to be built from scratch with careful consideration of the task at hand. In that sense, gravitational-wave detection epitomizes both the promise and the challenge of AI in physics: a tool that can enhance sensitivity and accelerate discovery, but only when deeply entangled with the infrastructure and ethos of the field.

When the objects of study are fundamental particles – the quarks, leptons, force mediating bosons from the Standard Model or beyond – the only option is to watch them interact. Particle colliders help do this by smashing particles together at near-light speeds: $E = mc^2$ in action. The faster and more energetic the collisions, the more new particles are created from the collision.

Particle physics experiments, particularly those at colliders, generate some of the largest and most homogeneous datasets in science. The homogeneity comes from control: every proton-proton collision occurs under nearly identical conditions, unlike the irreproducible complexity of most natural systems. But with such enormous volumes of data, even rare statistical fluctuations can start to look meaningful. An observation at 3σ significance, 99.7 percent confidence – which is as unlikely as flipping a coin and getting all heads in eleven tosses – will appear somewhere in billions of collisions simply by chance. To guard against mistaking these coincidences for discoveries, physicists set the bar at 5σ significance, 99.99994 percent confidence, before declaring a new particle or phenomenon to be real.

At the Large Hadron Collider, detectors see up to thirty million collisions per second. Storing everything is impossible, so experiments rely on a multilayered pipeline to select the most informative events. A hardware trigger, built from custom electronics including field-programmable gate arrays (FPGAs), selects about one in four thousand collisions within microseconds. Surviving events pass to software triggers running on vast CPU and GPU farms, which apply more detailed criteria. Only then are the remaining candidates reconstructed: raw detector hits assembled into tracks, tracks combined into jets, and objects tagged as electrons, photons, or heavy-flavor decays. Continuous calibration and alignment account for detector drift, while anomaly detection safeguards against both faulty signals and unexpected physics. To hedge against hidden biases, a small random subsample of untriggered events is also stored.

Machine learning has supported this pipeline for decades. Neural networks were already used for pattern recognition in the 1990s, long before “deep learning” became fashionable.¹² What has changed is the scale and specialization of the models. Today, ultrafast classifiers run directly on FPGAs, graph neural networks find tracks in dense collision environments, convolutional networks suppress pile-up noise, and autoencoders assist in calibration and monitoring.¹³ Many of these models are trained on simulated detector responses using tools such as GEANT4 that

provide the labeled data necessary for supervised learning. Because this introduces the risk of mismodeling, collaborations continually validate against control samples, retrain on real collisions, and parameterize known sources of uncertainty.¹⁴

Machine learning now underpins nearly every layer of the data pipeline, operating under some of the most stringent real-time requirements in science. In the first-tier trigger, custom models implemented on FPGAs or application-specific integrated circuits execute with latencies below fifty nanoseconds, filtering millions of events per second; higher-level triggers and offline analyses then refine those selections through increasingly complex inference.¹⁵ The result is one of the most elaborate and rigorously validated deployments of AI in any scientific domain.

Yet its success depends as much on culture as on computation. The LHC community prizes redundancy, interpretability, and continuous cross-checks, values that make it possible to trust automation at such scale. Far from replacing physics expertise, AI is absorbed into an infrastructure of verification honed over decades. The LHC epitomizes one archetype of physics data: immense, controlled, homogeneous, and sifted through industrial-scale analysis pipelines. Machine learning thrives here not by offering general-purpose models but by becoming a set of specialized tools tailored to the unforgiving statistical and cultural demands of particle physics.

On a scale larger than gravitational waves and particle collisions but still invisible to the naked eye lie atoms and the systems they form: molecules, crystals, proteins, nanoclusters. These systems are the playground of condensed matter physics, materials science, and chemistry, fields that vary more in the types of questions they're interested in than in their subject matter. Atomic systems illustrate a different face of physics, one rich in methods and degrees of control but fraught with inconsistency and approximation.

Atoms can, in principle, be arranged, synthesized, and probed in myriad ways. The atomistic sciences are overflowing with an alphabet soup of characterization techniques: X-ray diffraction (XRD) for crystallography, transmission electron microscopy (TEM) for atomic imaging, angle-resolved photoemission (ARPES) for electronic structures, nuclear magnetic resonance (NMR) for local environments, and many more. Some of these tools fit on a lab bench, others demand infrastructure the size of a city block. Synchrotrons accelerate electrons around kilometer-scale rings to generate intense X-rays. Free-electron lasers are so bright they vaporize the very sample they probe. Neutron sources offer complementary information by scattering particles without charge. Even magnets (the strongest sometimes drawing 10 percent of a city's power) become instruments to test matter under extreme conditions.¹⁶

This diversity is a blessing and a curse. There are plenty of data, but they are patchy, inconsistent, and often specialized. No two probes see exactly the same

thing, and experimental fluency is required to navigate across techniques. This patchwork of formats, conventions, and partial coverage makes it extremely difficult to build cohesive benchmarks

But if experiments are diverse, the underlying theory seems, at first glance, wonderfully unified. The Schrödinger equation, in principle, describes all quantum systems of electrons and nuclei. But in practice it can only be solved exactly for hydrogen. Even helium requires approximation. For anything larger (molecules, crystals, surfaces), physicists descend a hierarchy of methods: quantum Monte Carlo, coupled cluster, Hartree-Fock, and density functional theory (DFT). Each method trades off accuracy, interpretability, and computational cost.

The workhorse is DFT, which balances tractability with reasonable fidelity. It has enabled vast databases of calculated properties and powered high-throughput searches for new materials.¹⁷ Yet it is far from push-button. Every system demands choices of functionals, basis sets, and convergence criteria; no single setup works universally. The field relies on a diverse ecosystem of DFT codes. Ensuring they produce consistent results has required deliberate coordination: a major benchmarking effort a decade ago confirmed that, with care, independent implementations agree to within a few millielectronvolts per atom.¹⁸

Fortunately, DFT's approximations are largely systematic; standard methods underestimate band gaps and misrepresent some electron correlations. These biases become part of the interpretive culture of the field: researchers know which properties to trust and which to treat skeptically.

Machine learning enters atomistic science in two complementary ways. On the experimental side, AI is now used to interpret noisy spectra, accelerate image analysis in electron microscopy, and guide beamline experiments in real time.¹⁹ At large facilities, the deluge of data – petabytes from experiments running at synchrotrons, neutron sources, or free-electron lasers – makes automated analysis not just convenient but essential.²⁰

On the simulation side, neural networks trained on quantum mechanical calculations act as surrogates. “Machine-learned interatomic potentials” can approximate energies and forces with near-DFT accuracy but orders of magnitude faster, enabling molecular dynamics of systems thousands of times larger or longer-lived than conventional calculations allow. Graph and Euclidean neural nets embed physical symmetries directly into their architectures, ensuring that predictions of atomistic systems respect invariance under permutation or rotation, respectively.²¹ These constraints are not decorative: without them, the models are less reliable and often less efficient. Because symmetries pervade physics, equivariant models have also found use in other domains, including high-energy physics.²²

Notably, the training data for these surrogates are themselves shaped by approximation. A model built on DFT inherits DFT's biases. It may be fast and accurate on its terms, but is only as trustworthy as the assumptions encoded in its source.

This is why validation against experiments, or cross-checks with higher-level theory, remains indispensable.

Atomistic systems thus present a paradox. They offer tremendous experimental freedom, myriad ways to probe matter, and a unifying theoretical foundation in quantum mechanics. But in practice, the data are messy and the theory is approximated, leaving wide gaps for interpretation. Machine learning is valuable precisely because it can bridge across these mismatches: extending the reach and fidelity of simulations, denoising spectra, interpolating between scarce experiments. At the same time, it inherits the fragility of its inputs. In atomistic science, AI is not a magic key but a tool that thrives only when tethered tightly to physical insight and cross-validation.

The path forward is not about replacing the physicist's craft but making it easier to use and extend. The real opportunity lies in creating systems that let ideas, data, and methods travel more freely – between instruments and analysis pipelines, between subfields that rarely share language, and between the computational and experimental parts of the discipline. This need for connective infrastructure and shared abstractions echoes the conclusions of the recent NSF Workshop on the Future of Artificial Intelligence and the Mathematical and Physical Sciences, which emphasize integration across the full scientific cycle – from data acquisition to theory – rather than isolated algorithmic advances.²³

Progress often depends on small, local acts of translation: building the glue code that connects detectors to data streams, theory to simulation, or one community's conventions to another's. The next step is to make that translation easier and more durable. That means developing abstractions that encode physical intuition without demanding heroics, infrastructures that let existing setups plug into modern data workflows, and frameworks that make the growing body of scientific knowledge more searchable and interoperable. Only once those foundations are in place can we build models capable of higher abstraction, models that reason across scales and modalities, whose reliability rests on the very interconnectedness of the system. The challenge and the promise are the same: to make physics more continuous, cumulative, and playful by building AI systems that extend how we think and experiment.

At the moment, every functional application of AI in physics rests on deep, local expertise. These models succeed because physicists and computer scientists have embedded physical structure directly into them: preserving symmetries, constraining transformations, encoding the logic of experiment, and validating until skepticism is satisfied. This is not automation; it is modeling by another name.

But there is a cost to customization. Performant models are rarely “plug and play.” In atomistic modeling, for example, productive use requires fluency in mul-

multiple domains at once: training and validating neural networks, generating and interpreting quantum calculations such as DFT, running molecular dynamics, and connecting atomic trajectories to experimental observables. That combination of skills is uncommon, and without safe abstractions, every new model must be built nearly from scratch. The DFT code VASP succeeded not by replacing expertise but by embedding it – decades of tuning functionals and convergence criteria – so researchers could build on reliable defaults without mastering every detail. Machine learning in physics still lacks such durable interfaces.

Furthermore, there is a friction between custom architectures with useful abstractions and computation performance. Equivariant neural networks that encode symmetry through tensor products or group representations are a striking case in point. These methods are mathematically demanding and map poorly onto GPU hardware optimized for dense matrix multiplications. Equivariant and physics-informed architectures often run slower and scale worse than their generic transformer cousins, not because they are flawed but because far less engineering effort has gone into making them efficient.

Bridging this gap falls to researchers and students who sit between disciplines. They tune architectures, reconcile units and conventions, and write the glue code that lets symbolic physics talk to numerical data. Each “AI breakthrough” in physics hides a scaffolding of bespoke pipelines and late-night debugging sessions. These efforts produce remarkable results, but they are fragile. If a few key people move on, the knowledge of how to make things work can vanish.

Fragmentation compounds the problem. Physics is not one field but many, each with its own data culture, instruments, and timescales. In some domains, like particle physics and cosmology, massive homogeneous datasets demand industrial-scale pipelines. In others, like condensed-matter physics or quantum information, data are heterogeneous and deeply contextual. Even defining a benchmark is hard: every experiment has its own geometry, calibration, and noise characteristics, and analysis tools evolve alongside the instruments themselves. The meaning of any dataset is inseparable from how it was made.

The near-term challenge is to make this integration less punishing, to design environments where physics and machine learning can meet without so much heroic labor. That will mean building abstractions that carry domain knowledge forward so future students can explore new ideas rather than constantly rebuilding the same ones. Ultimately, success will look less like end-to-end automation and more like the emergence of robust, trustworthy interfaces between experiment, simulation, and learning.

Building tools that observe the world is difficult, slow, and expensive. The design, coordination, and construction needed for particle accelerators, telescopes, synchrotron beamlines, and new generations of microscopes

often take decades. Their purpose is to reach new regimes of measurement, making signals visible that were previously inaccessible. Progress in physics depends as much (if not more) on these instruments as on the theories that interpret them.

Artificial intelligence, by contrast, evolves on the timescale of months. New architectures, training schemes, and software frameworks appear faster than any experimental cycle could ever adapt. This creates a widening gap between how quickly AI develops and how slowly physical infrastructure can evolve. Cryogenic electron microscopy (cryo-EM) is a vivid example. The method has been in development since the 1970s, but only after advances in detector hardware and reconstruction algorithms did it achieve near-atomic resolution, earning a Nobel Prize in 2017.²⁴ Now, just as the hardware has matured, the computational stack beneath it is being rewritten again: motion-correction, denoising, and 3D-reconstruction algorithms are rapidly incorporating machine learning. Cryo-EM shows how long hardware takes to stabilize and how quickly the software layers atop it can shift, forcing scientists to rebuild analysis pipelines even as the microscopes themselves remain unchanged.

This mismatch recurs across physics. Instruments that take decades to build are paired with software that reinvents itself every few months. For now, the most meaningful role of AI is therefore to help us make better use of the tools we already have: to coordinate experiments, interpret results, and connect observations into larger bodies of knowledge.

Even discovering what data exist can be arduous: datasets are scattered across facilities and formats, and their interpretive context often lives only in lab notebooks or custom scripts. Depositing data or creating benchmarks requires documenting calibration, noise, and uncertainty in exhausting detail. AI could ease some of that burden by learning to recognize overlap among experiments, automate the collation of data streams, and link partial results into a coherent map of what has been observed. A system that can suggest the next most informative experiment or identify patterns across instruments would make scientific progress cumulative rather than episodic. In this sense, the frontier is not just faster computation but better coordination, building connective tissue that allows insights to travel between experiments, institutions, and scales.

Even our accumulated knowledge faces similar coordination problems. The scientific literature contains centuries of data, models, and insight, yet much of it is effectively dark matter, too vast and unevenly indexed for any individual to navigate. Large language models already hint at how this might change. As scientific literature assistants, they could help researchers trace relationships across subfields, summarize prior work, and connect results that use different vocabularies or conventions. The technology exists now; what is missing is reliability and integration. Modern systems still invent citations, blur context, and struggle with paywalls and proprietary archives. Building assistants that respect provenance, verify sources,

and work within the publishing ecosystem could make the literature itself a searchable, dynamic experiment wherein ideas are as accessible as data.

Even more transformative would be using these tools to translate across disciplinary boundaries. Every subfield of physics and every neighboring science carries its own language, priorities, and sense of what counts as progress. A discovery in one area may be routine in another; a null result in one domain may be a signal in a different framing. Large models trained with and guided by social scientists and historians of science could make these distinctions legible, mapping not just concepts but the cultures that produced them. Reinventing the wheel is the tragedy of the unsearchable; AI, used thoughtfully, could help us see where the wheels already exist, why they were built the way they were, and what else they could be used to move.

Once AI systems help integrate context and coordinate knowledge, they can begin to expand what kinds of physical systems we can represent and reason about. The goal is not simply to make existing workflows more efficient, but to reach the gray zones between our simplifying regimes, places where theory, experiment, and simulation have each struggled to stand alone.

Multiscale modeling is one of these frontiers. Physical theories and computational tools are typically built for specific domains: quantum mechanics for electrons, molecular dynamics for atoms, continuum mechanics for materials. Between them lie regimes where the spatial and temporal separations that justify standard approximations break down, and where small-scale structure and collective behavior become inseparably intertwined. Machine learning can help bridge these seams, learning mappings between levels of description or inferring effective parameters directly from data. In doing so, it can connect first-principles calculations to emergent phenomena without the loss of interpretability that often comes from coarse-graining.

Another challenge is multimodality. The same system can be observed in many incompatible ways, through diffraction, microscopy, spectroscopy, or transport, each yielding a partial and often biased picture. AI could help integrate these perspectives, learning how one type of measurement constrains another or how multiple views combine into a consistent description. Such models could even propose new directions for investigation, identifying which probes or conditions might most reduce uncertainty or reveal an unmeasured degree of freedom.

These capabilities mark a qualitative shift: from AI as a post-hoc analyst of data to AI as an active participant in the modeling process. A multiscale, multimodal model does not replace theory or experiment; it acts as connective tissue between them. It can represent phenomena too rich for any single equation yet too structured to be left to blind fitting.

Generative models already play a role in guiding experiments, but only in narrow, well-structured settings in which automation is feasible and objectives are

clearly defined. Models can propose candidate molecules, crystal structures, or experimental parameters, and such tools have begun to accelerate discovery in condensed matter physics, chemistry, and materials science. Yet these systems still operate within fixed contexts: they explore what is already measurable rather than helping us recognize what could be measured – phenomena or configurations we have not yet thought to test.

AI that can plan or orchestrate experiments is a recurring scientific dream, one that always feels five or ten years away. The difficulty is not only conceptual but physical. True experimental orchestration requires laboratories, instruments, and workflows designed for continuous machine-human interaction or for no human interaction at all, except when the machine inevitably breaks, because scientific instruments are never as robust as we would like them to be. Changing those environments is slow and capital-intensive, far harder than deploying new software. What sounds like a modest step, letting models choose the next experiment, is in fact one of the most ambitious because it touches the very infrastructure of how we do science.

The deeper promise is for AI that engages with experiment and simulation as a reasoning partner, able to weigh uncertainty, recognize when a model's assumptions fail, and suggest new hypotheses rather than merely new samples. Such tools would help researchers explore vast design spaces and build intuition about them. As these systems mature, AI's role in physics will evolve from tool to collaborator. Models that can reason across scales and modalities will not just analyze data but participate in theory-building, connecting disparate observations through shared structure. Most profoundly, they may help reframe questions themselves. By surfacing degeneracies, highlighting hidden assumptions, or uncovering structural parallels across systems, AI can extend physics' vocabulary, creating space for new abstractions and new ways to reason.

Realizing this vision will require trust, transparency, and shared intuition between people and models. The challenge is not simply to make AI accurate but to make its reasoning legible, ensuring that it participates in physics' culture of approximation, verification, and debate. Done well, this partnership could make physics more exploratory: a discipline whose models do not just describe the world but help us ask better questions about it.

Artificial intelligence is already woven into how we run experiments, simulate systems, and analyze data. The challenge ahead is to make these successes easier to sustain and build upon. The most valuable systems in the near term will amplify human judgment, helping us sift through torrents of data, accelerate simulations, and make instruments more responsive to real-time context. Their strength lies in preserving the link between data and the conditions under which those data were made.

The next stage will continue to require cultural fluency as much as technical progress. Physicists will need to learn the languages of learning; machine-learning researchers will need to respect the values that keep physics resilient to hype, including transparency, testability, and the habit of doubt. These dual responsibilities, linking technical, cultural, and infrastructural progress, mirror the themes identified in the NSF workshop's recent community paper, which calls for sustained collaboration between domain scientists and AI researchers to build trustworthy, interoperable systems.²⁵

It also means recognizing that shared infrastructure and composable models are as essential as headline-grabbing results. Progress will depend on quieter forms of work: building infrastructure that connects modeling with measurement, designing experiments that yield training signals as well as insights, and ensuring that physical context is carried forward rather than lost. The promise of AI in science is beyond prediction; it is understanding – revealing why systems behave as they do and how their patterns connect across scales.

In physics, understanding has always come from abstraction: finding the right variables, effective models, and questions to make the complex tractable. AI may expand this vocabulary. Architectures that encode invariances, workflows that reflect physical constraints, and interfaces that expose reasoning all extend what physicists can express and test.

Better tools help us ask better questions. Structure shapes answers and guides what we consider to be askable in the first place. Some questions fail because they rest on the wrong assumptions, as when prequantum physics imagined electrons orbiting nuclei or when relativity overturned the ether. Good tools can reveal those hidden assumptions and help us reframe inquiry itself.

If we get this right, AI will broaden our reach for what we can ask, discover, and know. It will help link ideas across subfields, lower the barriers between experiment, computation, and theory, and make room for richer, testable abstractions. The goal remains the same: to make the complexity of the universe knowable to the human mind.

AUTHOR'S NOTE

The author thanks Jane Halpern and Elyssa Hofgard; Professors Jesse Thaler, Phil Harris, and Sean Lubner; and Drs. Ryley McConkey and Emily Oliphant for their contributions to the structure and framing of this essay. The author also benefited from discussions with the members of the Atomic Architects research group and at the MIT Schwarzman College of Computing ML-Driven Discovery Retreat.

This work reflects insights gained through collaborations with advisors, colleagues, and students across physics and machine learning, and through research conducted at national laboratories and their user facilities, including Fermilab, CERN, Argonne National Laboratory, the National High Magnetic Field Laboratory, and Lawrence Berkeley National Laboratory. The author is a member of the NSF Institute for Artificial Intelligence and Fundamental Interactions (IAIFI) and the NSF Institute for Artificial Intelligence and Materials (AI-MI), and acknowledges the broader NSF AI program for its support of emerging roles for AI in physics.

ABOUT THE AUTHOR

Tess Smidt is Associate Professor of Electrical Engineering in the MIT Department of Electrical Engineering and Computer Science. She is also Principal Investigator of the Research Laboratory of Electronics and of the Atomic Architects research group. She has recently published in such journals as *Nature Communications*, *Advances in Neural Information Processing Systems*, and *Physical Review Research*.

ENDNOTES

- ¹ Inés Gil Botella and DUNE Collaboration, “The Deep Underground Neutrino Experiment (DUNE) Program,” presented at the 42nd International Conference on High Energy Physics (ICHEP2024), Prague, Czech Republic, July 18–24, 2024; and A. Weltman, Philip Bull, Stefano Camera, et al., “Fundamental Physics with the Square Kilometre Array,” *Publications of the Astronomical Society of Australia* 37 (2022), <https://doi.org/10.1017/pasa.2019.42>.
- ² Benjamin P. Abbott, Richard Abbott, Thomas D. Abbott, et al., “Observation of Gravitational Waves from a Binary Black Hole Merger,” *Physical Review Letters* 116 (2016): 061102.
- ³ Jack Dongarra and David Keyes, “The Co-Evolution of Computational Physics and High-Performance Computing,” *Nature Review Physics* 6 (2024): 621.
- ⁴ Laura Boggia, Carlos Cocha, Fotis Giasemis, et al., “Review of Machine Learning for Real-Time Analysis at the Large Hadron Collider Experiments ALICE, ATLAS, CMS and LHCb,” arXiv (2025), <https://doi.org/10.48550/arXiv.2506.14578>.
- ⁵ Abbott, Abbott, Abbott, et al., “Observation of Gravitational Waves from a Binary Black Hole Merger.”
- ⁶ Benjamin P. Abbott, Richard Abbott, Thomas D. Abbott, et al., “Characterization of Transient Noise in Advanced LIGO Relevant to Gravitational Wave Signal GW150914,” *Classical and Quantum Gravity* 33 (13) (2016): <http://doi.org/10.1088/0264-9381/33/13/134001>; and Anamaria Effler, Robert M. Schofield, Valery V. Frolov, et al., “Environmental Influences on the LIGO Gravitational Wave Detectors during the 6th Science Run,” *Classical and Quantum Gravity* 32 (3) (2015), <http://doi.org/10.1088/0264-9381/32/3/035017>.
- ⁷ Jonas Buchli, Brendan Tracey, Tomislav Andric, et al., “Improving Cosmological Reach of a Gravitational Wave Observatory Using Deep Loop Shaping,” *Science* 389 (6764) (2025), <https://doi.org/10.1126/science.adw1291>; and Chris Whittle, Ge Yang, Matthew Evans, and Lisa Barsotti, “Machine Learning for Quantum-Enhanced Gravitational-Wave Observatories,” *Physical Review D* 108 (2023), <https://doi.org/10.1103/PhysRevD.108.043034>.

- ⁸ Ethan Marx, William Benoit, Alec Gunny, et al., “Machine-Learning Pipeline for Real-Time Detection of Gravitational Waves from Compact Binary Coalescences,” *Physical Review D* 111 (2025), <https://doi.org/10.1103/PhysRevD.111.042010>.
- ⁹ Deep Chatterjee, Ethan Marx, William Benoit, et al., “Rapid Likelihood Free Inference of Compact Binary Coalescences Using Accelerated Hardware,” arXiv (2024), <https://doi.org/10.48550/arXiv.2407.19048>.
- ¹⁰ Gravitational-Wave Candidate Event Database (GraceDB), “S250830m,” submitted August 30, 2025, <https://gracedb.ligo.org/superevents/S250830m/view>.
- ¹¹ Mario Krenn, Yehonathan Drori, and Rana X. Adhikari, “Digital Discovery of Interferometric Gravitational Wave Detectors,” *Physical Review X* 15 (2025), <https://doi.org/10.1103/PhysRevX.15.021012>.
- ¹² Bruce Denby, “Neural Networks and Cellular Automata in Experimental High Energy Physics,” *Computer Physics Communications* 49 (3) (1988), [https://doi.org/10.1016/0010-4655\(88\)90004-5](https://doi.org/10.1016/0010-4655(88)90004-5); Leif Lönnblad, Carsten Peterson, and Thorsteinn Rögnvaldsson, “Finding Gluon Jets With a Neural Trigger,” *Physical Review Letters* 65 (1990): 1321, <https://doi.org/10.1103/PhysRevLett.65.1321>; Bruce Denby and Stephan L. Linn, “Spatial Pattern Recognition in a High Energy Particle Detector Using a Neural Network Algorithm,” *Computer Physics Communications* 56 (3) (1990): 293–297, [https://doi.org/10.1016/0010-4655\(90\)90015-S](https://doi.org/10.1016/0010-4655(90)90015-S); and Carsten Peterson, Thorsteinn Rögnvaldsson, and Leif Lönnblad, “JETNET 3.0—A Versatile Artificial Neural Network Package,” *Computer Physics Communications* 81 (1–2): 185–220, [https://doi.org/10.1016/0010-4655\(94\)90120-1](https://doi.org/10.1016/0010-4655(94)90120-1).
- ¹³ Boggia, Cocha, Giasemis, et al., “Review of Machine Learning for Real-Time Analysis at the Large Hadron Collider Experiments ALICE, ATLAS, CMS and LHCb”; and Allison M. Deiana, Nhan Tran, Joshua Agar, et al., “Applications and Techniques for Fast Machine Learning in Science,” arXiv (2021), <https://doi.org/10.48550/arXiv.2110.13041>.
- ¹⁴ Boggia, Cocha, Giasemis, et al., “Review of Machine Learning for Real-Time Analysis at the Large Hadron Collider Experiments ALICE, ATLAS, CMS and LHCb.”
- ¹⁵ Deiana, Tran, Agar, et al., “Applications and Techniques for Fast Machine Learning in Science.”
- ¹⁶ TaMaryn Waters, “It’s So Hot in Tallahassee, Even the Power Hungry FSU MagLab Needed to Take a Break,” *Tallahassee Democrat*, July 31, 2025, <https://www.tallahassee.com/story/news/2025/07/31/tallahassee-mag-lab-magnetic-power-bill-fsu-extreme-heat-experiment-florida/85445060007>; and National High Magnetic Field Laboratory, “By the Numbers,” <https://nationalmaglab.org/about-the-maglab/facts-figures/by-the-numbers> (accessed October 20, 2025).
- ¹⁷ Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, et al., “Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation,” *APL Materials* 1 (1) (2013), <https://doi.org/10.1063/1.4812323>; Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld, “Quantum Chemistry Structures and Properties of 134 Kilo Molecules,” *Scientific Data* 1 (2014): 140022, <https://www.nature.com/articles/sdata201422>; and Lowik Chanussot, Abhishek Das, Siddharth Goyal, et al., “Open Catalyst 2020 (OC20) Dataset and Community Challenges,” *ACS Catalysis* 11 (10) (2021), <https://pubs.acs.org/doi/10.1021/acscatal.0c04525>.
- ¹⁸ Kurt Lejaeghere, Gustav Bihlmayer, Torbjörn Björkman, et al., “Reproducibility in Density Functional Theory Calculations of Solids,” *Science* 351 (6280) (2016), <https://doi.org/10.1126/science.aad3000>.

- ¹⁹ Benediktus Madika, Aditi Saha, Chaeyul Kang, et al., “Artificial Intelligence for Materials Discovery, Development, and Optimization,” *ACS Nano* 19 (30) (2025): 27116–27158.
- ²⁰ Jonathan Carter, John Feddema, Doug Kothe, et al., *Advanced Research Directions on AI for Science, Energy, and Security: Report on Summer 2022 Workshops* (U.S. Department of Science, Office of Scientific and Technical Information, 2023), <https://www.anl.gov/sites/www/files/2023-06/AI4SESReport-2023-v5.pdf>.
- ²¹ Ryan Jacobs, Dane Morgan, Siamak Attarian, et al., “A Practical Guide to Machine Learning Interatomic Potentials—Status and Future,” *Current Opinion in Solid State Materials Science* 35 (2025), <https://doi.org/10.1016/j.cossms.2025.101214>.
- ²² Alexander Bogatskiy, Sanmay Ganguly, Thomas Kipf, et al., “Symmetry Group Equivariant Architectures for Physics,” paper submitted to the Proceedings of the U.S. Community Study on the Future of Particle Physics (Snowmass 2021), March 14, 2022, <https://arxiv.org/pdf/2203.06153>.
- ²³ Andrew Ferguson, Marisa LaFleur, Lars Ruthotto, et al., “The Future of Artificial Intelligence and the Mathematical and Physical Sciences (AI+MPS),” arXiv (2025), <http://arxiv.org/abs/2509.02661>.
- ²⁴ Yifan Cheng, Nikolaus Grigorieff, Pawel A. Penczek, and Thomas Walz, “A Primer to Single-Particle Cryo-Electron Microscopy,” *Cell* 161 (3) (2015); and Allison Doerr, “Cryo-Electron Tomography,” *Nature Methods* 14 (2017), <https://doi.org/10.1038/nmeth.4115>.
- ²⁵ Ferguson, LaFleur, Ruthotto, et al., “The Future of Artificial Intelligence and the Mathematical and Physical Sciences (AI+MPS).”

Thinking & Doing Science in the Age of AI

Alison Noble

We are in a new era of advancing science through AI. But the tremendous progress is not just about technology: as scientists have adopted AI-based technologies, they have changed how they do science, redefining the role of human intelligence and the human scientist in the scientific endeavor. This essay reflects on some of these changes and some of the current challenges and tensions that science and scientific communities are facing as we grapple with how best to work with AI to advance science and ensure society's continued trust in scientists and AI-based scientific evidence.

As an engineering scientist who trained in computer vision and has subsequently worked in medical image analysis for nearly thirty years, it has been a privilege to observe and experience firsthand how, in a remarkably short timeframe, artificial intelligence-based techniques have advanced an applied-science discipline. It is hard to believe that it was just over a decade ago (or three generations of UK PhD students) that the first computer-vision methods achieved (average) human-level performance at automated object recognition in one of the early computer-vision data challenges.¹ No longer just a curiosity of computer-vision researchers working at the interface with medical imaging and radiology, there are numerous examples today of automated and assistive AI-based methods for disease detection, anatomical segmentation, and disease characterization that have been translated from concept to clinical research tools and regulatory-approved medical diagnostics used in health care settings. Clinical AI has emerged as a very active area of academic clinical medicine, with its own dedicated journals and conferences; researchers study issues around accuracy, bias, safety, and explainability of deployed AI-based methods and their readiness to be adopted safely into workflows.

Transformations of this kind are mirrored in a number of data-driven science disciplines in which the pattern recognition capabilities of AI can match, or in some cases exceed, human-level capabilities. The 2024 Nobel Prize in Physics, awarded in recognition of artificial neural network methods, and in Chemistry, for using AI systems to create novel proteins and to solve the protein-folding prediction problem, represented a “coming of age” of AI tools for scientific discovery.² Significant AI-related advances are also being made in other data-rich areas of science, such

as large-scale climate and public-health predictive modeling, and in other areas where research communities are pooling together and releasing large-scale datasets, such as astrophysics.³ Elsewhere, AI is being introduced as a productivity tool to accelerate real-world workflows and streamline experimental laboratory processes and analyses. Examples range from robotic chemistry lab assistants that can, in theory, work around the clock, video analytics for ethology and conservation, and AI software coding tools for software development.⁴ In these cases, the speed at which the AI assistant works has been key to their adoption and success. Much of science prior to the current AI era has traditionally been associated with reflection, not speed. So is AI-based science now challenging the traditional model of scientific advancement, or is it just an ever-expanding suite of new tools that automate existing procedures and accelerate discoveries in existing experimental data? Should we worry that artificial intelligence will replace human intelligence and critical thinking in science, or embrace change and adapt? The answer to the latter question is probably a bit of both.

Because many AI systems require enormous volumes of training data, the earliest adopters of AI in science have, perhaps predictably, come from data-driven science fields that have ready sources of high-quality and well-curated large datasets and that have employed highly skilled computational scientists working alongside domain experts in a small team. Effective as this has been, it is not a scalable approach across the whole of science; there simply are not enough highly skilled computational scientists to meet demand. Training domain experts in AI is also not trivial, and approaches to teaching non-AI scientists essential AI skills are continuously evolving with the AI techniques themselves. Today, for instance, I would argue that a “science and AI” (SciAI) scientist does not necessarily need to be able to program a deep-learning architecture from scratch because of the increasing availability of open access AI models. But they do need to understand the principles of design, build, and testing for the AI-modeling techniques they are using, as well as how to place use in the context of the science. They also need to understand how to rigorously evaluate AI models and challenge their potentially erroneous outputs. This is where human-in-the-loop science domain knowledge and skill are still paramount.

Data, and in particular access to high-quality and well-curated large datasets, are essential to ensuring that AI models accurately and reliably represent the scientific problems they are designed to address.⁵ The complexities of data management, curation, and stewardship should not be underestimated; for many researchers, obtaining approval to share and access data – essential for building models and reproducibility – is currently one of the greatest barriers in SciAI research. Experts on data management and data governance are in high demand and become valued members of science and AI research teams. Good data governance is required to ensure that data used to train and test AI models are used responsibly and ethically.

Best practices in scientific data governance and management of data privacy and copyright lag behind AI algorithms and are complicated by regional, national, and sector-specific legal requirements. Data governance for medical research use may, for example, have to satisfy GDPR (General Data Protection Regulation) requirements in Europe as well as HIPAA (Health Insurance Portability and Accountability Act) rules in the United States. The legal, regulatory, and ethical issues associated with data governance are particularly challenging for global science collaborations, in which researchers need to navigate different national and international data regulations that are evolving at different rates.

Some science communities have for decades been working collaboratively on data curation and on construction of open-data challenges. The European Molecular Biology Laboratory European Bioinformatics Institute (EMBL-EBI) is a good example. In 1994, the EMBL-EBI set up the Critical Assessment of Protein Structure Prediction (CASP), an international competition for computational model prediction of protein folding, which was won by the AlphaFold2 AI system at CASP14. Meanwhile, organizations such as Hugging Face and Kaggle have provided platforms to enable research communities to collaborate and benchmark computational methods on open datasets. With sensitive data areas like health care, research is shifting from a data-sharing model, whereby raw data are transferred between partners typically covered by a data-sharing agreement, to a data-access model, which has tighter controls on data privacy and security, with data remaining at the source of data generation.

One data-access approach allows AI analysis to take place only in a secure data-analytics environment. Data-access and compute charges are needed to cover the running costs and sustainability of such research infrastructure, and debates on fairness and equity have been ongoing for some time, as fee models favor industry and other well-funded groups and are out of reach for many early-career researchers and researchers with limited resources.

For groups that wish to collaborate to train a single model but cannot share data between sites, federated learning may provide a solution. In this case, a (global) model is sent to participating clients (for instance, where each has a radiology imaging database), each client uses its local, sensitive data to train and improve the model locally, and the client sends updated model parameters, but not the data, back to the centralized system, which then aggregates updated model parameters for all clients to give the updated global model. This approach has been applied in COVID-19 analysis, for instance, but there have been few other real-world scenario examples of research use so far.⁶ This is because of outstanding theoretical challenges, including how to effectively model tasks for the “small community experiment” of a small number of clients and highly heterogeneous, possibly multimodal data; and practicalities, including computational resource requirements at client sites. Open-source tools, such as Flower, that provide the infrastructure

for implementing federated learning–based modeling are gaining traction at the current time.⁷

Finally, some research areas (such as computer vision and large language models) have traditionally used large training datasets generated by (sometimes indiscriminately) scraping the internet for data. This sourcing raises questions about data copyright. But there is currently no international consensus on how to deal with copyrighted data used to train AI models for scientific research, highlighting the complexities and tensions between the intellectual property rights of the data author, open science, and the advancement of science by AI.

It is important to embrace the positives that the new AI tools and capabilities offer for scientific breakthrough, innovation, and potential economic and social impact. But we must also recognize that the use of AI tools in science challenges some of the established norms of scientific research – and that research ethics has never been more important in science than today.

Responsible AI (RAI) is central to trustworthy, safe, and ethical development and use of AI-based scientific methods. Community-led initiatives, including a collaboration between the National Academy of Sciences, the Annenberg Public Policy Center of the University of Pennsylvania, and the Annenberg Foundation Trust at Sunnyslans, and work by UNESCO and Responsible AI UK have set out to identify AI-expert consensus in this area and to encourage adoption of RAI principles and best practices.⁸

Scientific rigor and transparency matter. They are foundational principles of science as an evidence-based discipline. All AI models will make errors, typically due to limitations of the training data (such as data distribution bias, data size, or gaps in representational completeness of a problem) and the model itself (observed as parameter over- or underfitting). Why should we worry about this? After all, these are not new issues in scientific data modeling, and there are ways to identify and in some cases mitigate these factors. However, the opaqueness (black-box nature) of AI models makes it easier for researchers to accept an AI-generated conclusion when it is wrong, particularly if it gives a result that matches their hypothesis (the so-called echo-chamber effect in critical thinking). The current trend of preprint publication of scientific research without peer-review of the article is not helping either. The extent to which this is leading to false claims and publication of work that cannot be reproduced is worrying scientific communities. In the context of AI, good scientific practice requires researchers to rigorously design, implement, and test a model to ensure reproducibility, and to rigorously analyze bias, accuracy, safety, and explainability (BASE). Peer reviewers of articles featuring AI models cannot verify the models by reproducing them, so they depend instead on trusting the authors' descriptions of methods and (where appropriate) benchmarking with other published models on common data. Yet surprisingly few science

journals or conferences currently insist on rigorous evaluation and reporting, typically just requiring reporting on AI model accuracy and (sometimes) explainability. Beyond the components of BASE, ideally all scientific AI work should be reported according to FAIR (findable, accessible, interoperable, reusable) principles, as well as cover descriptions of datasets and encourage code and data to be made open for other researchers to use for benchmarking, where possible.

To address these concerns, comprehensive consensus-based guidelines and standards of reporting have been proposed in some disciplines, including CLAIM (AI in medical imaging) and REFORMS (general machine learning–based science).⁹ Adoption of such checklists is relatively new; it is important that these initiatives are given high visibility and community-wide support to improve standards of reproducibility and transparency, including by embedding them in the training of SciAI scientists.

The degree to which open science principles have been adopted in different scientific disciplines to date varies significantly. Communities that have traditionally worked together on data curation and data and code sharing have found this transition easier; many work together on data challenges. Those who work with sensitive real-world data that typically have restricted access, as with AI in medical imaging, follow a mixed model of working with public and private datasets. This creates a tension between researchers who insist that new method assessment should include analysis of public data and researchers who model novel tasks with novel private data, for which a public data equivalent does not exist. Following FAIR principles is hardest for disciplines that have traditionally comprised siloed groups, whose datasets are their assets and competitive edge. Time will tell if the challenges of reproducibility in AI and the desire to use AI to answer new scientific questions in such areas will shift attitudes and behavior from competition to partnership.

As with reproducibility, recent advances in AI, and particularly when generative AI (GenAI) is used as part of the research method, are requiring new thinking about the ways we look at scientific integrity. GenAI can generate synthetic scientific data, including images and text, and perform tasks such as text summarization, automated code generation, and image editing. As AI-generated (synthetic) content becomes more realistic and more prevalent, it must not be confused with real-world observations. In science, GenAI in its various guises, including large language models (LLMs), offers many advantages, ranging from literature synthesis to experimental design to writing, but it needs to be used with care. For this reason, an interdisciplinary panel convened by the National Academy of Sciences proposed five principles of human accountability and responsibility when using GenAI in science: 1) transparent disclosure and attribution, 2) verification of AI-generated content and analyses, 3) documentation of AI-generated data, 4) ethics and equity, and 5) continuous monitoring, oversight, and public engagement.¹⁰ The need for ethical

standards in AI-assisted writing and greater transparency in LLM use is also discussed, for instance, in recent work by bioethics scholar Sebastian Porsdam Mann and colleagues.¹¹ Issues of attribution also extend to the patenting of inventions involving AI. Patents are typically awarded for human creativity and currently there is no clear guidance on how to manage co-inventions between humans and AI.

No discussion of AI in science can ignore how AI is changing the future of scientific publishing. Throughout the history of science, scientists have communicated to their peers and the public in continuously changing ways, from peer-to-peer letter writing and public lectures in the early days to today's potpourri of dissemination methods for what are often global audiences, including peer-reviewed articles, databases and code repositories, conferences, and podcasts and web videos, to name a few. The scientific publishing industry is itself undergoing significant changes to incorporate AI and automate processes, but the discussion here will focus on perspectives related to the scientist (as the customer) and on how LLMs – the special case of generative AI that focuses on language-related tasks and is adept at search and mimicking human writing – are impacting science article content and article peer review.

The emergence of LLMs that can perform routine search and summarization/content discoverability at remarkable speed and scale means the end of literature and systematic reviews as we know them today. Many researchers take this as good news because it frees up their time for more creative tasks. Others argue that while current LLM-based summarization tools can capture an article's story, findings, and impact, they cannot assess the trustworthiness of code, methods, results, and analysis as well as people. If so, then humans still have roles in checking the AI assistant's results and performing original synthesis – though that may not be for long, depending on the literature reviewing task and how the technology develops. This is also an area where emerging human-AI collaboration technologies such as learning-to-defer, in which AI performs a task but defers to a human when the AI is unsure, may play a useful role.¹²

For peer review, the challenge and opportunity are to figure out how best to share the burden between AI and humans while ensuring the quality of the review process. There is already a wide spectrum of ways to publish scientific articles, from the preprint model of publishing first and discussing later to traditional peer-review of articles submitted to journals and conferences. These approaches serve different purposes in science (to share scientific findings with others, to create a scientific record, to gain professional recognition and influence) and we should respect and celebrate this diversity. However, it is well acknowledged that in many science disciplines, the current journal peer-review system is broken: the volume of papers submitted and demand for review exceeds human scholar availability. Additionally, human peer review can be prone to bias and can struggle to detect fraudulent work. Hence there is significant interest from publishers and scientists in understanding

what role LLMs can play (or not) in supporting the review process and which types of scientific work LLMs can assess. Scientists have different views on this. Moreover, as science journalist Miryam Naddaf has noted, it is critical to set clear standards of transparency so both the authors submitting articles for review and readers of the journal are informed that AI is part of the review process.¹³

If, in this new era of science and AI, our notions of scientists and the scientific method are changing, then what is the role of human intelligence and the human scientist?

While there are recent papers presenting early ideas and demonstrations of AI systems (agentic and nonagentic) that can automate entire scientific discovery processes, the AI tools we will increasingly see in AI-based science for the foreseeable future are assistive-AI systems (narrow AI) tuned to the needs, and built on the data, of a specific science domain.¹⁴

If AI systems can meet the (high) expectations of scientists and can be accepted as a team member in a science laboratory, does this change our expectations of the role and skill set of human scientists? These changes relate to how humans and AI can effectively work together, and who does what in the thinking and doing of science.

Where AI fully automates processes, such as in an experimental laboratory, jobs for humans will undoubtedly disappear. Elsewhere, the breadth of skills required by a human SciAI scientist are increasingly becoming an interdisciplinary mix of science domain expertise and knowledge, AI, and research ethics. For instance, researchers using AI tools in classic experimental disciplines such as biology, physics, and chemistry may now need to master data (analytical) skills, blurring the distinction between a computational scientist and experimental scientist of days gone by. Many AI tools are AI assistants, so the human scientist's role includes challenging the AI to verify outputs.

A decade ago in my own research area, medical image analysis, we would have labeled medical imaging AI technologists working with clinicians as interdisciplinary. Today, we also work with experts in human factors and with experimental psychologists to understand not only the accuracy and reliability of the systems we develop but to study, for instance, how human trust is essential for end user acceptance and safe deployment of those systems.¹⁵ This means the subject is becoming truly interdisciplinary, combining computer science, engineering, medicine, and the social sciences.

Do individuals necessarily need an interdisciplinary mix of skills? Probably not. And there are lessons academia might learn from the private sector in terms of how to effectively conduct SciAI research and development, by which a team of computer scientists, high-performance computing engineers, data managers and data governance specialists, science domain experts, and social scientists co-design, -build, and -evaluate models. Building scientific teams of this disciplinary

breadth and depth comes at a cost, but where the potential scientific gain is great, this model of interdisciplinary scientific collaboration, including public-private sector partnerships, is likely to become more common.

As a final point, I want to comment on recent research on the effect of AI on critical thinking. Critical thinking is a fundamental cognitive skill and an essential part of problem-solving and reflective thinking in scientific work. A recent study by AI and society scholar Michael Gerlich found a significant negative correlation between frequent AI tool use and critical thinking.¹⁶ This suggests that as we move forward with more AI integration into the scientific method, we need to ensure that the benefits of cognitive offloading and the desire to be first in AI-based scientific discovery are not at the expense of critical thinking. If we do not, human scientists risk losing the habit of thinking. While AI tools may assist with the doing, the thinking in science and AI needs to remain the responsibility of the human.

Artificial intelligence technologies are supporting scientists in many disciplines to advance scientific discovery and improve productivity. Our responsibility as scientists is to ensure rigor in our standards of reproducibility of scientific evidence, the scientific integrity of AI-based work, and that human critical thinking stays at the heart of science advancement. The rapid adoption of new AI technologies for task automation and human-AI collaboration to empower human scientists is all but inevitable. While scientists are rightly concerned about how AI technologies work and behave, understanding the human element of human-AI collaboration in SciAI research may be the greater challenge.

ABOUT THE AUTHOR

Alison Noble is the Technikos Professor of Biomedical Engineering at the University of Oxford. She is also Vice President and Foreign Secretary of the Royal Society and Chair of the Royal Society working group that produced the policy report *Science in the Age of AI* (2024). She previously served as Director of the Institute of Biomedical Engineering and as an Associate Head of Division at the University of Oxford. She has recently published in such journals as *Nature Biomedical Engineering*, *npj Digital Medicine*, *Medical Image Analysis*, and *Proceedings of the AAAI Conference on Artificial Intelligence*.

ENDNOTES

- ¹ Olga Russakovsky, Jia Deng, Hao Su, et al., “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision* 115 (3) (2015): 211–252.
- ² John Jumper, Richard Evans, Alexander Pritzel, et al., “Highly Accurate Protein Structure Prediction with AlphaFold,” *Nature* 596 (7873) (2021): 583–589, <https://doi.org/10.1038/s41586-021-03819-2>.
- ³ Multimodal Universe Collaboration (Eirini Angeloudi, Jeroen Audenaert, Micah Bowles, et al.), “The Multimodal Universe: Enabling Large-Scale Machine Learning with 100TB of Astronomical Scientific Data,” arXiv (2024), <https://doi.org/10.48550/arXiv.2412.02527>.
- ⁴ Tianwei Dai, Sriram Vijayakrishnan, Filip T. Szczypiński, et al., “Autonomous Mobile Robots for Exploratory Synthetic Chemistry,” *Nature* 635 (8040) (2024): 890–897, <https://doi.org/10.1038/s41586-024-08173-7>; and Max Bain, Arsha Nagrani, Daniel Schofield, et al., “Automated Audiovisual Behavior Recognition in Wild Primates,” *Science Advances* 7 (46) (2021), <https://doi.org/10.1126/sciadv.abi4883>.
- ⁵ National Academy of Sciences, *Toward a New Era of Data Sharing: Summary of the U.S.-UK Scientific Forum on Researcher Access to Data* (National Academies Press, 2024), <https://doi.org/10.17226/27520>.
- ⁶ Ittai Dayan, Holger R. Roth, Aoxiao Zhong, et al., “Federated Learning for Predicting Clinical Outcomes in Patients with COVID-19,” *Nature Medicine* 27 (10) (2021), <https://doi.org/10.1038/s41591-021-01506-3>.
- ⁷ See Flower AI, <https://flower.ai>.
- ⁸ Wolfgang Blau, Vinton G. Cerf, Juan Enriquez, et al., “Protecting Scientific Integrity in an Age of Generative AI,” *Proceedings of the National Academy of Sciences* 121 (22) (2024), <https://doi.org/10.1073/pnas.2407886121>.
- ⁹ Ali S. Tejani, Michail E. Klontzas, Anthony A. Gatti, et al., “Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update,” *Radiology Artificial Intelligence* 6 (4) (2024), <https://doi.org/10.1148/ryai.240300>; and Sayash Kapoor, Emily M. Cantrell, Kenny Peng, et al., “REFORMS: Consensus-Based Recommendations for Machine-Learning-Based Science,” *Science Advances* 10 (18) (2024), <https://doi.org/10.1126/sciadv.adk3452>.
- ¹⁰ Blau, Cerf, Enriquez, et al., “Protecting Scientific Integrity in an Age of Generative AI.”
- ¹¹ Sebastian Porsdam Mann, Anuraag A. Vazirani, Mateo Aboy, et al., “Guidelines for Ethical Use and Acknowledgement of Large Language Models in Academic Writing,” *Nature Machine Intelligence* 6 (2024): 1272–1274, <https://doi.org/10.1038/s42256-024-00922-7>.
- ¹² Hussein Mozannar and David Sontag, “Consistent Estimators for Learning to Defer to an Expert,” arXiv (rev. 2021), <https://arxiv.org/abs/2006.01862>.
- ¹³ Miryam Naddaf, “AI Is Transforming Peer Review—and Many Scientists Are Worried,” *Nature*, March 26, 2025, <https://www.nature.com/articles/d41586-025-00894-7>.
- ¹⁴ Chris Lu, Cong Lu, Robert Tjarko Lange, et al., “The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery,” arXiv (2024), <https://doi.org/10.48550/arxiv.2408.06292>; and Yoshua Bengio, Michael Cohen, Damiano Fornasiere, et al., “Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?” arXiv (2025), <https://doi.org/10.48550/arxiv.2502.15657>.

- ¹⁵ Angus Nicolson, Elizabeth Bradburn, Yarin Gal, et al., “The Human Factor in Explainable Artificial Intelligence: Clinician Variability in Trust, Reliance, and Performance,” *npj Digital Medicine* 8 (2025): 658, <https://doi.org/10.1038/s41746-025-02023-0>.
- ¹⁶ Michael Gerlich, “AI Tools in Society: Cognitive Offloading and the Future of Critical Thinking,” *Societies* 15 (1) (2025): 6, <https://doi.org/10.3390/soc15010006>. See also the references therein.

Field Theory: AI as Social Science Question, Object & Tool

Alondra Nelson

Uses of advanced artificial intelligence are changing how societies organize labor, govern, produce knowledge, and make meaning. In light of these developments, this essay argues that AI models, tools, and systems pose three interrelated imperatives for social science: they demand renewed attention to social theories of how technology, human experience, and social order are entangled; they require study as objects of inquiry in their own right; and they offer capabilities that may transform – or upend – the practice of social investigation itself. From Weber’s analysis of rationalization to Du Bois’s study of technology and inequality to contemporary scholarship on algorithmic governance, the essay examines what social science distinctively offers: the capacity to historicize the apparently unprecedented, to trace connections across scales, and to center those most affected by technological change. It identifies how algorithmic systems are remaking the distribution of opportunity and risk as a central task of social inquiry and asks what futures social science might help bring into being.

The emergence of advanced artificial intelligence – from generative models that produce text, sound, and images, to the expanding array of algorithmic tools now embedded across social life, to deep learning and agentic systems that may transform research – has become a touchstone for contemporary debates about society. The introduction of AI models, tools, and systems is both a marker of human resourcefulness and a portent of societal disruption. These technologies arrive with promises to enhance human capacities, accelerate scientific discovery, and generate new cultural forms. Yet their use also threatens to erode human autonomy and displace human judgment, and is heightening inequality, destabilizing political systems, and obscuring accountability.

For social science, this moment is urgent. The introduction of AI confronts researchers with at least three interlinked imperatives: to renew explanations for how technological change is entangled with human experience and social order; to examine AI as an object of inquiry; and to assess AI as a methodological tool, the speed and scope of which may change, but also distort, the practice of social research. These imperatives, in combination, define what it means to study society in the age of algorithms and automated power.

This is not the first time social science has addressed technology as important to the analysis of society. Since their inception, the social sciences have been bound up with the theorization of technological change in areas including economic life, industrial production, labor relations, and education. Max Weber traced how technical rationalization and bureaucracy remade authority. W. E. B. Du Bois revealed how industrialization intersected with economic and racial inequality. Later, Herbert Marcuse exposed how technology could reshape freedom and thought itself. Their work exemplifies how moments of technological upheaval have spurred the development of foundational ideas in the social sciences.

Yet advanced AI poses challenges that may push this tradition to its limits. Unlike industrial machines that extended human labor or bureaucratic forms that rationalized tasks, AI tools operate through rapid statistical patterning and large-scale data processing that may defy easy comparison to prior technologies. Their deployment can unsettle categories that once seemed stable – work, knowledge, governance, even personhood – and in doing so, demand normative and conceptual innovation from social scientists. The sociotechnical phenomena that earlier theorists illuminated reappear today but refracted through algorithmic systems of scale, opacity, and reach that may strain inherited frameworks.

What is at stake is more than retrospective understanding. If social scientists wait until AI systems are fully entrenched, we will be left to document their consequences after the fact. The opportunity instead is to shape trajectories in parallel time to articulate critical frameworks, surface risks, generate concepts and evidence that guide public choices about design and governance, and advance research that heightens our understanding of contemporary society.

Social scientists confronting artificial intelligence join a tradition of grappling with novel technological developments. We have long recognized that technologies are not merely neutral instruments but forces that redistribute power and shape meaning. Classic theories of society were often forged in response to technologically mediated social disruption: the upheavals of the Industrial Revolution, the displacements of empire, the reorganization of work and politics after global war. Each wave of transformation demanded not only empirical description but also sense-making. Artificial intelligence, with its promises and perils, presents a similarly paradigmatic moment, one that calls for retrieval of these theoretical traditions and creative extensions of them.

Among the earliest to grasp these dynamics was Max Weber, who sought to make sense of how industrial capitalism reorganized authority and legitimacy. In *The Protestant Ethic and the Spirit of Capitalism* (1905), he observes that “this order is now bound to the technical and economic conditions of machine production which . . . determine the lives of all individuals . . . with irresistible force.”¹ For Weber, the factory system was not just an economic structure but a relentless drive to organize

life through calculation, efficiency, and control. He describes the result as an “iron cage”: a social order in which bureaucratic and technical logics, once unleashed, confined human beings within systems of rationalization and constrained their freedom. This iron cage imagery resonates powerfully today. Algorithmic management tracks the movements of Amazon warehouse workers, the routes of Uber drivers, and call center productivity; it automates employment decisions and imposes a web of continuous surveillance. As factories in the Industrial Revolution regimented time and labor, AI systems now impose regimes of quantification across domains from warehouse logistics to knowledge work.² Weber helps us see that these uses of artificial intelligence are not simply a bundle of techniques but a system of order – an institutional logic that reorganizes human relationships under the banner of rationalization.

Yet Weber’s focus on rationalization left crucial social dimensions unexamined. At the same moment, W. E. B. Du Bois insisted that technological change could not be understood apart from relations of power, particularly those structured by race and empire. In *The Negro Artisan* (1902), he documents how automation displaced Black craftspeople, eroding forms of autonomy and skill that had been essential to their civic standing. Without political protections, he argues, technological change deepened inequality rather than alleviating it.³ Later, in *Darkwater: Voices from Within the Veil* (1920), Du Bois extends this critique to the global stage. He observes that “Western science” and industry had perfected machines of production without developing mechanisms for just distribution, turning technical command into a tool of domination. Du Bois insists that technologies absent political safeguards are not emancipatory but instead reinforced hierarchies of race, class, and empire.⁴ His analysis points us toward a central question for the use of advanced AI today: Who benefits, who is displaced, and how are inequalities reinscribed through the deployment of these new systems?

By the mid-twentieth century, Herbert Marcuse had given this critique a new emphasis, showing how technological rationality could suppress critical reflection and reinforce conformity. In *One-Dimensional Man* (1964), he argues that the modern exercise of power no longer depended primarily on overt coercion but worked through technology itself. Efficiency and productivity became unquestioned ends, narrowing thought and desire and foreclosing the imagination of alternatives.⁵ This dynamic is legible in today’s AI systems. Recommendation engines steer attention and desire, curating environments that make certain possibilities appear natural while obscuring others. Large language models produce persuasive outputs that blur the boundary between information and manipulation; as with Marcuse’s diagnosis, constraints on human agency seem inevitable precisely because they arrive in the guise of technical systems. His work reminds us that the power of technology lies not only in what it enables, but in how it constrains the horizons of thought and action.

Taken together, Weber, Du Bois, and Marcuse remind us that technologies do not merely augment human capacity; they can shuffle social life, redistribute power, and enframe meaning. Returning to them matters now not because their theories are perfectly fitted to the AI era, but because they teach us to ask the right kinds of questions. Their insights remain indispensable for interpreting AI, even as new conditions require us to build upon them with creativity and care. To retrieve this tradition is therefore not to rehearse it mechanically but to renew it, elaborating new categories and methods appropriate to this moment.

Artificial intelligence introduces forms of complexity and uncertainty that call for new conceptual tools. Machine-learning systems, for example, can generate outputs or identify patterns that were not programmed in advance and that may be unanticipated by developers. Equally significant is the reach of AI technologies across domains. Unlike a machine tethered to a single site, like a hospital or factory, AI systems are pervasive, spanning multiple social spheres. A credit score shapes not only financial access but also housing and employment opportunities.⁶ Content-moderation algorithms influence both everyday expression and national politics.⁷ Supply-chain optimization changes corporate logistics and reverberates through global labor markets and geopolitics.⁸ These interconnections reveal AI tools and models as systems with multiplier effects that may ripple beyond their originally intended applications or their immediate sites of use.⁹

To this unpredictability and cross-domain reach, add a third defining feature of advanced AI: opacity – not necessarily intrinsic to the tools themselves, but the outcome of commercial decisions to deploy systems that are not fully understood and rarely transparent. This, combined with proprietary claims, the rush of the so-called AI race, and lack of regulatory oversight, means that AI systems can be difficult to interrogate. Opacity, in this fuller sense, is not merely a technical condition but a political and economic strategy: systems are rendered difficult to interrogate not because complexity makes transparency impossible, but because commercial incentives and institutional arrangements make opacity profitable. This is precisely what makes AI a central object of social science inquiry rather than a narrow technical concern.¹⁰

These features – unpredictability, cross-domain reach, and opacity – also recast enduring questions of power. The rationalization of life, the entrenchment of inequality, and the narrowing of thought all persist in the AI era but refracted through systems and institutions that require reimagining from social scientists. How pervasive AI use organizes power and constrains imagination must be analyzed with attention to the scale, speed, and infrastructural reach that earlier theorists did not foresee.

Scholars across the social sciences have begun this work of theoretical renewal. In sociology, accounts of the “ordinal society” update classical analyses of rational-

ization, showing how algorithmic ranking and scoring intensify social stratification and redefine social membership.¹¹ In political science, research on democratic deliberation examines the conditions under which algorithmic tools might support rather than supplant collective decision-making – and also cautions how the use of artificial intelligence threatens democracy, including representation, accountability, and trust.¹² Economists have argued that accounting for interactions between researchers and machine learning tools can help refine theory generation, and have developed task-based frameworks demonstrating that the effects of AI on labor and inequality are not technologically determined but depend on choices about how these systems are designed and deployed.¹³ Theories of digital labor have shown how data-labeling centers and microwork platforms deploy rules of discipline, supervision, and fragmentation that differently reproduce and deepen, rather than transcend, industrial regimes of control.¹⁴ Studies of race and technology have revealed how algorithmic systems presented as neutral actually encode and reproduce racial hierarchy, extending historical patterns of discrimination into automated decision-making.¹⁵

Yet even powerful emerging technologies do not determine their own applications or outcomes. Research on algorithmic biotechnologies has underscored their “social life” and how communities contest and repurpose scientific tools, offering conceptual models for understanding how the meanings and uses of AI remain sites of struggle rather than settled outcomes.¹⁶ Research on automated financial trading has extended sociologist Erving Goffman’s concept of the “interaction order” to algorithmic systems to explain how trading algorithms queue, compete, and deceive through electronic order books, generating emergent dynamics that no individual actor designed or fully controls.¹⁷ Such work suggests the theoretical innovation required of this moment: drawing on disciplinary traditions while forging insights adequate to the distinct qualities and challenges of advanced AI.

But the full theoretical architecture for understanding AI in society remains to be built. What is needed is not simply a catalog of its effects across sectors, but integrated accounts of how algorithmic systems are transforming the fundamental coordinates of social existence: the nature of work and economic value, the character of public life and political authority, the production and circulation of knowledge, and the very boundaries of human agency. Such frameworks will grasp artificial intelligence not as a discrete technology but as an infrastructural layer that increasingly mediates how institutions operate, how individuals are seen and sorted, and how collective futures are imagined and foreclosed.

This moment requires social science to both retrieve its critical traditions and enlarge them – to think beyond the categories that earlier transformations made salient and to craft new concepts commensurate with present uncertainties. Building this account will require sustained collaboration across disciplines and a willingness to conduct empirical research while theorizing at the scale of

the AI transformation. But renewal cannot be achieved in abstraction. It requires engagement with the models, tools, and systems through which artificial intelligence is being deployed in society. Linking critical traditions with emerging research, social science should approach the AI turn as demanding not only theoretical attention but empirical investigation.

For the social sciences, AI has become unavoidable. It touches every site of inquiry – political, economic, cultural, and epistemic. Therefore, to study it is not merely to analyze a suite of technologies, but to examine the “full AI stack” – from the extraction of critical minerals, to the energy demands of data centers, to the involvement of workers in development and adoption, to the design and circulation of AI products.¹⁸ Taken seriously, this means treating AI models, tools, and systems not as a technical enterprise but as a social one – the foundation for a social science of AI.

Across the social sciences, technologies are analyzed not as mere instruments but as human products that steer flows of resources, capital, and imaginaries, and whose design, ownership, circulation, and use are embedded in wider contexts.¹⁹ This perspective is essential for understanding AI systems that are built in laboratories and corporations but quickly exceed those sites, becoming embedded in society. They embody assumptions about what is valuable or efficient, shaping the distribution of visibility, recognition, and opportunity.²⁰ The integration of AI into social life therefore demands attention not only from scholars of the impacts of technology but from all who study the organization of society.

This becomes clear when examining how AI intersects with social relations. What appears as a media “feed” or “trending topic” is in fact the outcome of algorithmic decisions about visibility and engagement.²¹ Research has traced how recognition, intimacy, and status are increasingly mediated by AI systems, altering the conditions under which social ties are formed and maintained.²² Surveillance, too, has been remade: no longer simply recording behavior, algorithmic surveillance actively conditions it, nudging individuals toward certain choices, constructing probabilistic profiles that may preempt action altogether, and foreclosing privacy, all across entire populations.²³ AI thus emerges not only as an infrastructure of communication but as a mediating force – one that curates what is seen, conditions what is done, and influences how identity and social capital are constituted through computational logics operating simultaneously at intimate and aggregate levels.

A social science of AI can also interrogate labor markets, demonstrating that artificial intelligence does not simply substitute for or augment human work but reorganizes the relations of expertise and value through which economies are conventionally structured. Platforms such as Google, Amazon, and ByteDance deploy AI to coordinate production and labor, extracting value through data, scale, and algorithmic curation. These dynamics strain economic models premised on ratio-

nal actors: AI systems generate patterns of interaction between human behavior and the speed and scope of algorithms that traditional models may not capture.²⁴ Economist David Autor argues that this reorganization does not simply eliminate jobs but “reshapes expertise,” enabling more workers to take on responsibilities once limited to highly credentialed professionals.²⁵ Whether this strengthens or erodes middle-class livelihoods, however, depends on institutional design and policy.²⁶ Autor refines this point with research scientist Neil Thompson by showing that automation affects wages and employment differently depending on whether “expert” or “inexpert” tasks are displaced.²⁷ Analyzing millions of online job postings, economist Daron Acemoglu and colleagues trace how AI adoption alters skill demands, finding that it reduces vacancies outside of roles that are AI-intensive even as aggregate employment effects remain modest.²⁸ These labor and inequality effects are inseparable from the political and institutional contexts in which these technologies are deployed.

A social science of AI must therefore also investigate governance in the algorithmic age: how power is exercised, how legitimacy is claimed, and how publics can hold AI-mediated institutions to account.²⁹ Governments around the world now deploy algorithmic tools to allocate welfare benefits, regulate immigration, and police populations.³⁰ Sarah Brayne’s ethnography of the Los Angeles Police Department illustrates how so-called predictive policing systems reorganize discretion, expanding surveillance of marginalized communities while insulating the users of such technology from accountability.³¹ Sociologists therefore argue that social science’s engagement with artificial intelligence must include questions of governance and sustained attention to how AI produces inequality and concentrates power.³² Political scientists have explored whether it might support rather than supplant collective decision-making, examining how algorithmic tools could facilitate deliberation at scale while acknowledging the risks such technologies pose to public acceptance and participation.³³ These studies demonstrate that AI deployment reflects policy choices and organizational arrangements as well as technological affordances.

AI also demands attention as a cultural phenomenon – one that shapes not only what people see and hear but how meaning itself is produced. Recommendation algorithms curate the films, music, and news that billions of people encounter daily, structuring attention and taste.³⁴ These systems do not merely reflect existing preferences, they actively shape them, creating feedback loops between user behavior and algorithmic curation that make it difficult to disentangle choice from conditioning. Ethnographic research – including of cultural consumption and gig work – has traced how workers, artists, and content creators navigate these logics and how AI systems cultivate particular forms of subjectivity.³⁵ At the same time, these technologies are altering the infrastructure of public discourse, raising questions about the future of shared knowledge and trust.³⁶ Through these processes,

the deployment of AI alters the cultural ground on which social life unfolds, not determining outcomes but changing the terms on which sense is made.

The implications of AI use extend beyond governance and the law to cultural understandings of humanity itself. In particular, anthropologists and sociologists have explored how people interact with AI companions, assistants, and avatars in ways that blur the boundaries between tool, partner, and persona – raising concerns about emotional dependence and mental health.³⁷ Sociologist Marion Fourcade and political scientist Henry Farrell contend that large language models are already destabilizing everyday rituals that underpin trust in institutions.³⁸ Recommendation letters, performance evaluations, and peer review rely on repetition and formality as much as originality; when such rituals are generated by AI, their meaning shifts. Efficiency may increase, but the risk is an erosion of belief in the sincerity of others' words – a foundation on which social trust depends. Anthropologist Nick Seaver adds to this line of research by showing how the concept of "attention" has migrated from human psychology into the architecture of neural networks.³⁹ In machine learning, attention functions as a technical mechanism for weighing relevance, but in public life, it also names a fragile human faculty vulnerable to capture. These studies suggest that uses of AI not only extend human capacities but reconstitute them, turning attention and ritual into contested sites of agency and significance.

Finally, AI use poses challenges for knowledge production itself, including that produced by social science. Researchers increasingly rely on machine learning to analyze large datasets, identify patterns, and generate hypotheses at speeds no human team could match. These tools also introduce new epistemic risks. Anthropologist Lisa Messeri and cognitive scientist M. J. Crockett argue that AI is taken up in research as the Oracle, the Surrogate, the Quant, or the Arbiter – roles that lend an aura of authority and create "illusions of understanding," in which narrow outputs are misinterpreted as broad comprehension.⁴⁰ This raises questions about both the validity of AI-assisted findings and the distribution of epistemic authority: Who is positioned to evaluate claims generated by systems whose operations may be opaque? As AI tools become embedded in research workflows, social science disciplines face a reflexive challenge to study shifts in knowledge, trust, and epistemology in society while remaining attentive to how their own methods may be reoriented in the process.

At a high level, a consistent finding emerges from this research: AI is not an external force acting upon society from outside but an internal phenomenon constituted through social processes, institutional choices, market dynamics, and political struggles. Its effects are neither uniform nor preordained. They vary with context, design, and the distribution of power that determines how these technologies are built, deployed, and contested. A social science of AI must therefore be both empirical and normative, attentive to how algorithmic systems operate in

practice and to the values that should guide their development. The research surveyed here represents the beginnings of such an enterprise, one that treats AI not as an exogenous shock but as a central site where fundamental questions about inequality, authority, and human well-being are being negotiated.

The questions that remain are as consequential as any the social sciences have faced. Under what conditions do algorithmic systems deepen stratification, and under what conditions might they be designed to reduce it? How do interactions mediated by artificial intelligence reshape the formation of identity, community, and collective action? What institutional and political arrangements can preserve meaningful human agency in decision-making processes increasingly delegated to machines? What forms of governance are appropriate and effective for increasingly dynamic sociotechnical systems? How will societies reimagine, negotiate, or redraw the boundaries between human and machine? And as these AI models, tools, and systems are developed to be more capable, autonomous, and ubiquitous, what societal futures are being foreclosed or made possible by the choices embedded in their development? These are fundamental questions about social order and the distribution of benefits and burdens.

If AI technologies are a necessary focus of social inquiry, they are also increasingly vehicles through which social science itself is conducted. Just as the introduction of the survey, the census, statistical modeling, and qualitative data analysis platforms changed research practice in earlier periods, AI tools now extend the methodological repertoire of the social sciences in ways both generative and unsettling.⁴¹ The challenge is to discern how these tools might enable new ways of knowing without constraining inquiry to what is most easily computable or without allowing research agendas to be determined by readily available technology rather than important questions.

AI technologies may offer significant benefits for social research. By rapidly processing vast corpora of text, images, and data, they enable social scientists to analyze patterns that would have been impossible to detect with traditional methods. AI tools can help scholars extend analysis from the microlevel of individual interactions to the macrolevel of global networks, opening new vantage points on how culture, politics, and economy evolve. They may help accelerate research workflows, facilitate comparative analysis across time and place, and generate synthetic datasets that allow testing of counterfactuals. These capacities could advance disciplinary debates and strengthen the public role of social science, potentially equipping researchers to respond to urgent issues with rigor and dispatch, including about the social and political implications of AI technologies and the companies developing them.⁴²

Over a decade ago, I joined colleagues in calling for a new “computational social science.” In the journal *Science*, we argued that the massive digital traces generated by

commercial online activity opened unprecedented opportunities for social inquiry.⁴³ Harnessing these data required collaboration across disciplines – computer science, sociology, political science, economics, and law – and demanded careful attention to ethics, privacy, and interpretation. Computational social science promised to extend the scope of what social scientists could do and know, allowing contemporaneous analysis of social dynamics at the population level. But it also carried risks of distortion if pursued uncritically, treating data as ends in themselves rather than as a partial means of illuminating social life.⁴⁴

That call is now renewed in the age of AI, though with greater caution. Where computational social science once relied on analyzing digital traces – search queries, social media posts, mobile phone records – AI introduces generative and “predictive” systems that actively structure those traces and can even generate synthetic ones. Social scientists are no longer just analyzing human and social behavior as mediated by digital tools; they are also analyzing environments already shaped by artificial intelligence and increasingly using algorithms themselves to conduct research.⁴⁵ AI is not simply another source of data. It is an array of novel research instruments, sometimes productive and sometimes misleading, that reconfigures facets of social research and demands far more critical scrutiny than it has received to date.

Examples from across disciplines illustrate the potential and the limits of this methodological transformation. In sociolinguistics, large language models have opened new possibilities for analyzing language acquisition and processing. Trained on massive text corpora, these models learn broad statistical regularities of language – what linguist Emily M. Bender and colleagues have described as “stochastic parroting” – that researchers can nonetheless leverage to explore linguistic patterns.⁴⁶ While these models do not “understand” language as humans do, their performance provides a new lens for testing hypotheses about the structure of language and the nature of learning it.

In applied economics, machine-learning methods enable real-time analysis of market sentiment, drawing on streams of text from news, social media, and financial disclosures.⁴⁷ Economists can now study the feedback loops between information and behavior, revealing dynamics that were once opaque.⁴⁸ Yet these methods also raise concerns. Forecasting models that integrate AI tools may improve the accuracy of “predictions” at the cost of interpretability, and with limited capacity to explain the mechanisms driving research findings or economic outcomes. The field confronts a trade-off between robustness, methodological transparency, and shared evidentiary standards.

In sociology, the use of artificial intelligence can enable massive-scale content analysis that can trace cultural change or the diffusion of social activity across millions of documents and posts. Thus, researchers can now combine qualitative findings with computational breadth. For example, sociologist Christopher Bail’s experiments with AI-mediated social simulation demonstrate how algorithms can

model the dynamics of polarization, offering new ways to test hypotheses about how beliefs spread and harden.⁴⁹ These models open possibilities for exploring counterfactual explanations, but they also underscore the risk of naturalizing assumptions embedded in algorithmic design.

Political science has likewise embraced AI for modeling, particularly in the study of electoral behavior and policy outcomes. Models trained on historical data can attempt to forecast voter turnout, partisan shifts, or policy adoption.⁵⁰ Yet “predictive” success does not translate into explanatory depth. Algorithmic models can reproduce observed patterns without shedding light on underlying mechanisms, raising questions about how much weight these analyses should merit for explanation and theory building.

At the same time, AI-powered survey research offers efficiency and scalability but raises concerns about systematic exclusion and rigor. These methods may misrepresent vulnerable populations who are less digitally visible or whose linguistic patterns diverge from dominant training data, replicating patterns of exclusion that have long characterized both social research and the institutions it studies.⁵¹ Communities already subject to disproportionate surveillance – through policing, welfare administration, or immigration enforcement – may find themselves rendered simultaneously hypervisible to state systems and invisible to the research intended to illuminate their conditions. A study might generate findings that appear generalizable while omitting those most affected by policy.⁵² Moreover, the nascent use of AI-generated research participants may pose an “existential threat” to social science, as it destabilizes the boundary between human and artificial actors and calls into question whether online research can still reliably capture authentic human behavior.⁵³ These problems are not unique to the use of algorithmic tools in research, but the velocity of AI-assisted research amplifies their stakes.

The prudent use of AI technologies raises the possibility of exploring complex questions that have been central to social science but remain difficult to examine fully with existing approaches. How does collective action evolve and how is it sustained over time? How do informal economies operate at scale and how are they shaped by patterns of social interaction and influence? How does language influence cognition and culture? How do communities build resilience not only in the face of environmental or economic shocks but also amid shifting social and technological conditions? How do information, narratives, and norms shape decision-making, conflict, and cooperation across societies – both domestically and transnationally? How do social institutions sustain legitimacy and facilitate cooperation in rapidly changing social environments?

Realizing this potential depends on how AI tools are designed and integrated into social inquiry. Progress in this area will require technologies that do not yet exist or remain underdeveloped: models designed for interpretability and explain-

ability; systems trained on ethically acquired and representative data rather than convenience samples of the digitally visible or digitally vulnerable; tools capable of integrating qualitative and quantitative evidence and engaging interpretive complexity rather than collapsing it into discrete classifications. If AI developers in industry and academia oriented their work toward these goals – in partnership with social researchers and communities – the result could be not merely incremental improvements but a genuine expansion of what the social sciences can know and do.

Yet realizing this potential requires methods that artificial intelligence alone cannot provide. Ethnographic practice – long-term engagement, trust building, and attentiveness to what is not said as much as what is – can reveal the subtle ways AI is embedded in social life.⁵⁴ Such approaches are particularly essential for understanding how AI is experienced differently across communities shaped by race, class, and geography – differences that aggregate data routinely obscure. Without such grounding, studies risk reproducing the very opacity they critique. If ethnographic grounding is essential, so too is direct engagement with the systems under study. In work with journalist Julia Angwin and technologist Rina Palta, I examined how five widely used chatbots responded to questions about U.S. elections. The study found that these systems routinely generated inaccurate, incomplete, or misleading answers.⁵⁵ The project highlighted both the utility of AI in stress testing automated information systems and the necessity of independent, human-driven evaluation to assess their societal risks.

The lesson across these examples is that AI technologies can be uneven research tools. As Messeri and Crockett caution, AI risks becoming a “hammer seeking nails,” tempting researchers to fit questions to available tools rather than developing tools adequate for meaningful questions.⁵⁶ This dynamic threatens to narrow inquiry to what is most legible to algorithms while neglecting issues that require interpretive or contextual approaches. A related but distinct concern applies to the widespread use of off-the-shelf AI models. Trained on opaque, unrepresentative, or historically biased corpora, such tools compromise the validity and reproducibility of research even when the questions themselves are well conceived.

Meeting these challenges will require movement on several fronts. The most promising path forward lies in research that combines ethically developed and responsibly deployed algorithmic tools with the depth of qualitative and theoretical analysis. Machine learning might identify patterns across large datasets, which social scientists then interpret through ethnographic or historical methods that situate findings in social relations and illuminate their significance. Such collaborations could open forms of inquiry that neither algorithmic nor interpretive approaches can achieve alone, expanding the reach of social science while preserving a commitment to better understanding social complexity. This will demand vigilance, however, and resistance to the temptation to treat AI tools and systems as value-free instruments or to let computational outputs set the terms of social inquiry.

New validation techniques are also required. Many AI models, tools, and systems are black boxes. Incorporating them uncritically into research could render social science itself opaque. Social scientists need methods to benchmark AI-assisted findings against established approaches, to test robustness across models, and to make transparent the assumptions embedded in training data.⁵⁷ Without such validation, the field may cede its explanatory force to systems it cannot fully interrogate, producing results without the accountability and transparency on which scholarly credibility depends.

These problems are not incidental to proprietary AI models – they are fundamental to their design. When corporations control the most powerful systems, academic researchers become users rather than architects of their own methods. Training data remain undisclosed, model updates arrive without warning, and access can be revoked or priced beyond reach. Open-source alternatives – models, tools, and systems with training data, architectures, and governance that are transparent to research communities and others – address these epistemic limitations directly. There is a strong case that social science should refuse proprietary AI models entirely. The epistemic requirements of scholarly inquiry may be fundamentally at odds with profit-driven, black-boxed systems whose training data, architecture, and governance remain inaccessible to most researchers. Developing such alternatives requires investment and collaboration, but the payoff is research infrastructure that serves scholarship rather than market imperatives. This is not antitechnology sentiment but a defense of what distinguishes research from consultancy: the imperative that claims about the social world be contestable and answerable to communities of inquiry.

The development of renewed ethical frameworks for advanced AI use in social science is equally urgent and should draw on the field's insights about technology, power, and social order. Just as research standards for consent and protection of human subjects evolved over the twentieth century, so too must frameworks for AI-augmented research evolve now. This includes attention to how training data are sourced, how outputs are used, and how participants' autonomy is respected.⁵⁸ It also requires scrutiny of how AI-mediated research may reinforce the very inequities many social researchers study: for instance, by drawing on data generated through surveillance of marginalized communities, by contributing to environmental pollution, or by disrupting social cohesion. AI models, tools, and systems have both great potential and profound limitations. Used uncritically for social science, they risk narrowing the horizons of inquiry; used reflexively, they may expand how social science can contribute to public understanding. The aim is to develop a practice of social science research with artificial intelligence that is theoretically innovative, methodologically rigorous, ethically responsible, and socially relevant. Such a practice could enable researchers to address pressing social problems with new precision – work developed in partnership with affected communities, atten-

tive to historical inequities, and subject to ongoing evaluation. This future social science could not only illuminate social challenges but help envision and test interventions with both technical capacity and moral imagination.

Artificial intelligence is more than just another tool of modern life. It reorganizes labor, governance, knowledge, and culture, unsettling categories that once seemed stable, and compelling the social sciences to confront perennial questions of freedom, inequality, and meaning anew. The specifics will change; the models that dominate today will be superseded, the policy debates will shift, the applications we now find novel will become infrastructural and invisible. But the underlying challenge – of understanding how technologies remake the conditions of social existence – will persist, as will the need to remain attentive to who benefits, who is harmed, and what possibilities are opened and foreclosed.

History offers some guidance. Moments of rupture – industrialization, mass migration, the rise of digital networks – have always demanded conceptual reinvention. Social scientists of earlier generations developed new frameworks to make sense of factory labor, bureaucratic organization, and the transformation of public life by mass media. The advent of advanced AI represents another such rupture, distinguished by its reach, velocity, and entanglement with the most intimate dimensions of everyday life. AI now shapes the conditions under which perception, communication, and decision unfold, and does so in ways that strain our existing vocabularies.

For social science, this is not merely a challenge but a responsibility. Few other domains of inquiry are better positioned to address the issues that matter most today and in the future. These questions cannot be left to technologists alone. They demand the distinctive capacities of social science: the ability to trace connections between individual lives and structural forces, to historicize the apparently unprecedented, to center the perspectives of those most affected, and to insist that the way things are is not the way they must be.

The stakes could not be more consequential. AI is already shaping who gets hired, who gets surveilled, who receives care, and whose voice is amplified or suppressed. Left to the vagaries of market imperatives and technical optimization, these systems will reproduce and entrench the uneven social conditions that social scientists have long documented and sought to resolve. Yet this outcome is not inevitable.

This essay has been cautionary throughout. It must, however, conclude with some optimism. If social science rises to meet this moment, a new generation of researchers, trained in both AI-mediated methods and interpretive traditions, will produce knowledge that is at once rigorous and humane. Their work informs the design, deployment, and governance of AI systems that expand human well-being rather than rationing it by algorithm. Decision-makers draw on social scientific evidence to craft frameworks that are adaptive, accountable, and attentive to the communities most affected by technological change. The social science field of

play – perhaps no longer organized by disciplines as we know them – will forge new partnerships with developers and engineers who welcome normative scrutiny and legal guardrails, with compensated artists and humanists who bring different ways of framing and imagining, with communities who help shape research agendas, and with publics who demand that artificial intelligence serve the common good rather than narrow interests. Social science itself is transformed: its methods enriched by tools it has learned to use critically, its questions sharpened by the weight of what is at stake, and its influence amplified by recognition that serious, publicly engaged social theory and inquiry are indispensable to any future worth wanting.

AUTHOR'S NOTE

I am grateful to James Manyika and M. J. Crockett, as well as members of the Science, Technology, and Social Values Lab at the Institute for Advanced Study, for their thoughtful feedback on this essay.

ABOUT THE AUTHOR

Alondra Nelson, a Member of the American Academy since 2020, is the Harold F. Linder Chair and Professor in the School of Social Science at the Institute for Advanced Study. Between 2021 and 2023, she served as Deputy Assistant to President Joe Biden and as Acting Director and Principal Deputy Director for Science and Society in the White House Office of Science and Technology Policy. She was formerly the fourteenth President and CEO of the Social Science Research Council and the inaugural Dean of Social Science at Columbia University. She is the author of *The Social Life of DNA: Race, Reparations, and Reconciliation After the Genome* (2016), *Genetics and the Unsettled Past: The Collision of DNA, Race, and History* (2012, with Keith Wailoo and Catherine Lee), *Body and Soul: The Black Panther Party and the Fight Against Medical Discrimination* (2011), and *Technicolor: Race, Technology, and Everyday Life* (2001, with Thuy Linh Tu). Her most recent book, *Auditing AI*, coauthored with the Marquand House Collective, was published with MIT Press in April 2026.

ENDNOTES

- ¹ Max Weber, *The Protestant Ethic and the Spirit of Capitalism* (Routledge, 2001), 123.
- ² Lucas G. Pinheiro, "Protocols of Production: The Absent Factories of Digital Capitalism," *American Political Science Review* 119 (3) (2025): 1085–1098, <https://doi.org/10.1017/S003055424000911>.
- ³ W. E. B. Du Bois, *The Negro Artisan* (Atlanta University Press, 1902).
- ⁴ W. E. B. Du Bois, *Darkwater: Voices from Within the Veil* (Harcourt, Brace and Howe, 1920).
- ⁵ Herbert Marcuse, *One-Dimensional Man* (Beacon Press, 1964).
- ⁶ Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact," *California Law Review* 104 (3) (2016): 671–732; and Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown, 2016).
- ⁷ Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press, 2018); and Zeynep Tufekci, *Twitter and Tear Gas: The Power and Fragility of Networked Protest* (Yale University Press, 2017).
- ⁸ Tanya Carayannis, Naima Kane, Marie-Therese Png, and Alondra Nelson, *Summary Report: Workshop on the Geopolitics of Critical Minerals and the AI Supply Chain* (Science, Technology, and Social Values Lab, Institute for Advanced Study, 2025), https://www.ias.edu/sites/default/files/Critical-Minerals-Workshop_Summary-Report.pdf.
- ⁹ Alondra Nelson, *The Social Life of DNA: Race, Reparations, and Reconciliation after the Genome* (Beacon Press, 2016).
- ¹⁰ Alondra Nelson, "Algorithmic Agnotology: On AI, Knowledge, and Power," invited lecture presented at the 39th Annual AAAI Conference on Artificial Intelligence, Philadelphia, Pennsylvania, March 1, 2025.
- ¹¹ Marion Fourcade and Kieran Healy, *The Ordinal Society* (Harvard University Press, 2024).
- ¹² Hélène Landemore, *Open Democracy: Reinventing Popular Rule for the Twenty-First Century* (Princeton University Press, 2020); Sendhil Mullainathan, "Economics in the Age of Algorithms," *AEA Papers and Proceedings* 115 (2025): 1–23, <https://doi.org/10.1257/pandp.20251118>; Sarah Kreps and Doug Kriner, "How AI Threatens Democracy," *Journal of Democracy* 34 (4) (2023): 122–131, <https://www.journalofdemocracy.org/articles/how-ai-threatens-democracy>; and Mark Coeckelbergh, "Artificial Intelligence, the Common Good, and the Democratic Deficit in AI Governance," *AI and Ethics* 5 (4) (2025): 1491–1497.
- ¹³ Daron Acemoglu, "The Simple Macroeconomics of AI," NBER Working Paper 32487 (National Bureau of Economic Research, 2024), <https://doi.org/10.3386/w32487>; and Mullainathan, "Economics in the Age of Algorithms."
- ¹⁴ Pinheiro, "Protocols of Production."
- ¹⁵ Ruha Benjamin, *Race After Technology: Abolitionist Tools for the New Jim Code* (Polity, 2019).
- ¹⁶ Arjun Appadurai, *The Social Life of Things: Commodities in Cultural Perspective* (Cambridge University Press, 2014); Nelson, *The Social Life of DNA*; and Aaron Panofsky and Joan Donovan, "Genetic Ancestry Testing Among White Nationalists: From Identity Repair to Citizen Science," *Social Studies of Science* 49 (5) (2019): 653–681.
- ¹⁷ Donald MacKenzie, "How Algorithms Interact: Goffman's 'Interaction Order' in Automated Trading," *Theory, Culture & Society* 36 (2) (2019): 39–59.

- ¹⁸ Carayannis, Kane, Png, and Nelson, *Summary Report*.
- ¹⁹ Sheila Jasanoff, *Designs on Nature: Science and Democracy in Europe and the United States* (Princeton University Press, 2005).
- ²⁰ Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (NYU Press, 2018); Benjamin, *Race After Technology*; and Sarah Brayne, *Predict and Surveil: Data, Discretion, and the Future of Policing* (Oxford University Press, 2020).
- ²¹ Tufekci, *Twitter and Tear Gas*.
- ²² Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St. Martin's Press, 2018); and Fourcade and Healy, *The Ordinal Society*.
- ²³ David Lyon and David Murakami Wood, eds., *Big Data Surveillance and Security Intelligence* (University of British Columbia Press, 2021).
- ²⁴ Diane Coyle, *Markets, State, and People: Economics for Public Policy* (Princeton University Press, 2021).
- ²⁵ David Autor, "Applying AI to Rebuild Middle Class Jobs," NBER Working Paper 32140 (National Bureau of Economic Research, 2024), <https://doi.org/10.3386/w32140>.
- ²⁶ Erik Brynjolfsson, "The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence," *Dædalus* 151 (2) (Spring 2022): 272–287, https://doi.org/10.1162/daed_a_01915.
- ²⁷ David Autor and Neil Thompson, "Expertise," NBER Working Paper 33941 (National Bureau of Economic Research, 2024), <https://doi.org/10.3386/w33941>.
- ²⁸ Daron Acemoglu, David Autor, Jonathon Hazell, and Pascual Restrepo, "Artificial Intelligence and Jobs: Evidence from Online Vacancies," *Journal of Labor Economics* 40 (S1) (2022).
- ²⁹ Cynthia Dwork and Martha Minow, "Distrust of Artificial Intelligence: Sources & Responses from Computer Science & Law," *Dædalus* 151 (2) (Spring 2022): 309–321, https://doi.org/10.1162/daed_a_01918.
- ³⁰ Eubanks, *Automating Inequality*; Karen Levy, Kyla E. Chasalow, and Sarah Riley, "Algorithms and Decision-Making in the Public Sector," *Annual Review of Law and Social Science* 17 (2021): 309–334, <https://doi.org/10.1146/annurev-lawsocsci-041221-023808>; and Helen Margetts, "Rethinking AI for Good Governance," *Dædalus* 151 (2) (Spring 2022): 360–371, https://doi.org/10.1162/daed_a_01922.
- ³¹ Brayne, *Predict and Surveil*.
- ³² Tina Law and Leslie McCall, "Artificial Intelligence Policymaking: An Agenda for Sociological Research," *Socius: Sociological Research for a Dynamic World* 10 (2024).
- ³³ Hélène Landemore, "Can AI Bring Deliberation to the Masses?" in *Routledge Handbook of Collective Intelligence for Democracy and Governance*, ed. Stephen Boucher, Carina Antonia Hallin, and Lex Paulson (Routledge, 2024); and Henry Farrell and Hahrie Han, "AI and Democratic Publics," Knight First Amendment Institute at Columbia University, August 1, 2025, <https://perma.cc/6YU3-MPM6>.
- ³⁴ Nick Seaver, "Attention Is All You Need," in *Scenes of Attention: Essays on Mind, Time, and the Senses*, ed. D. Graham Burnett and Justin E. H. Smith (Columbia University Press, 2023), <https://doi.org/10.7312/burn21118>.
- ³⁵ Ibid.; and Lindsey Cameron, "Making Out While Driving: Relational and Efficiency Games in the Gig Economy," *Administrative Science Quarterly* 67 (4) (2022): 976–1013.
- ³⁶ Tufekci, *Twitter and Tear Gas*; and Gillespie, *Custodians of the Internet*.

- ³⁷ Linnea I. Laestadius, Andrea Bishop, Michael Gonzalez, et al., “Too Human and Not Human Enough: A Grounded Theory Analysis of Mental Health Harms from Emotional Dependence on the Social Chatbot Replika,” *New Media & Society* 26 (5) (2022): 2749–2767; and Celeste Campos-Castillo, Xuan Kang, and Linnea I. Laestadius, “Perspectives on How Sociology Can Advance Theorizing about Human-Chatbot Interaction and Developing Chatbots for Social Good,” arXiv (2025), <https://doi.org/10.48550/arXiv.2507.05030>.
- ³⁸ Marion Fourcade and Henry Farrell, “Large Language Models Will Upend Human Rituals,” *The Economist*, September 4, 2024, <https://www.economist.com/by-invitation/2024/09/04/large-language-models-will-upend-human-rituals>.
- ³⁹ Seaver, “Attention Is All You Need.”
- ⁴⁰ Lisa Messeri and M. J. Crockett, “Artificial Intelligence and Illusions of Understanding in Scientific Research,” *Nature* 627 (8002) (2024): 49–58, <https://doi.org/10.1038/s41586-024-07146-0>.
- ⁴¹ Jenny L. Davis and Mona Sloane, “AI’s Sociological Era,” *Social Science Computer Review*, November 7, 2025, <https://doi.org/10.1177/08944393251394690>; and Cathleen O’Grady, “AI May Upend Online Studies Critical to Social Science,” *Science*, December 19, 2025, <https://www.science.org/content/article/ai-may-upend-online-studies-critical-social-science>.
- ⁴² Rick Harrison, “AI Accelerates Research Across Social Sciences: Insights from Yale’s DISSC,” Institution for Social and Policy Studies (ISPS), Yale University, December 17, 2024, <https://isps.yale.edu/news/blog/2024/12/ai-accelerates-research-across-social-sciences-insights-from-yales-dissc>; and Igor Grossman, “Beyond the Hype: How AI Could Change the Game for Social Science Research,” *The Conversation*, July 3, 2023, <https://theconversation.com/beyond-the-hype-how-ai-could-change-the-game-for-social-science-research-208086>.
- ⁴³ David Lazer, Alex Pentland, Lada Adamic, et al., “Computational Social Science,” *Science* 323 (5915) (2009): 721–723.
- ⁴⁴ danah boyd and Kate Crawford, “Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon,” *Information, Communication & Society* 15 (5) (2012): 662–679, <https://doi.org/10.1080/1369118X.2012.678878>; Matthew Zook, Solon Barocas, danah boyd, et al., “Ten Simple Rules for Responsible Big Data Research,” *PLOS Computational Biology* 13 (3) (2017), <https://doi.org/10.1371/journal.pcbi.1005399>; and Diane Coyle, “Socializing Data,” *Dædalus* 151 (2) (Spring 2022): 348–359, https://doi.org/10.1162/daed_a_01921.
- ⁴⁵ J. Nathan Matias, “Humans and Algorithms Work Together—So Study Them Together,” *Nature* 617 (7960) (2023): 248–251, <https://doi.org/10.1038/d41586-023-01521-z>.
- ⁴⁶ Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🐦” in *FaccT ’21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2021), 610–623, <https://doi.org/10.1145/3442188.3445922>.
- ⁴⁷ Paul C. Tetlock, Maytal Saar-Tsechansky, and Sofus Macskassy, “More Than Words: Quantifying Language to Measure Firms’ Fundamentals,” *Journal of Finance* 63 (3) (2008): 1437–1467; Jon Ellingsen, Vegard H. Larsen, and Leif Anders Thorsrud, “News Media versus FRED-MD for Macroeconomic Forecasting,” *Journal of Applied Econometrics* 37 (1) (2022): 63–81, <https://doi.org/10.1002/jae.2859>; Matthew Gentzkow, Bryan T. Kelly, and Matt Taddy, “Text as Data,” NBER Working Paper 23276 (National Bureau of Economic Research, 2017), <https://doi.org/10.3386/w23276>; and Michele Costola, Oliver

- Hinz, Michael Nofer, and Lorian Pelizzon, "Machine Learning Sentiment Analysis, COVID-19 News and Stock Market Reactions," *Research in International Business and Finance* 64 (2023): 101881, <https://doi.org/10.1016/j.ribaf.2023.101881>.
- ⁴⁸ Coyle, *Markets, State, and People*; and Mullainathan, "Economics in the Age of Algorithms."
- ⁴⁹ Christopher A. Bail, "Can Generative AI Improve Social Science?" *Proceedings of the National Academy of Sciences* 121 (21) (2024), <https://doi.org/10.1073/pnas.2314021121>.
- ⁵⁰ Andrew Gelman and Gary King, "Enhancing Democracy Through Legislative Redistricting," *American Political Science Review* 88 (3) (1994): 541–559; and Ryan D. Enos, Anthony Fowler, and Lynn Vavreck, "Increasing Inequality: The Effect of GOTV Mobilization on the Composition of the Electorate," *Journal of Politics* 76 (1) (2014): 273–288.
- ⁵¹ Thomas Davidson, "Start Generating: Harnessing Generative Artificial Intelligence for Sociological Research," *Socius: Sociological Research for a Dynamic World* 10 (2024), <https://doi.org/10.1177/23780231241259651>; and Benjamin, *Race After Technology*.
- ⁵² M. J. Crockett and Laura Messeri, "AI Surrogates and Illusions of Generalizability in Cognitive Science," *Trends in Cognitive Sciences*, October 21, 2025, <https://doi.org/10.1016/j.tics.2025.09.012>; and Angelina Wang, Jamie Morgenstern, and John P. Dickerson, "Large Language Models That Replace Human Participants Can Harmfully Misportray and Flatten Identity Groups," *Nature Machine Intelligence* 7 (2025): 400–411.
- ⁵³ Sean J. Westwood, "The Potential Existential Threat of Large Language Models to Online Survey Research," *Proceedings of the National Academy of Sciences* 122 (47) (2025): e2518075122, <https://doi.org/10.1073/pnas.2518075122>.
- ⁵⁴ Roanne van Voosrt and Tanja Ahlin, "Key Points for an Ethnography of AI: An Approach Towards Crucial Data," *Humanities and Social Sciences Communications* 11 (2024): 337, <https://doi.org/10.1057/s41599-024-02854-4>.
- ⁵⁵ Julia Angwin, Alondra Nelson, and Rina Palta, *Report: Seeking Reliable Election Information? Don't Trust AI* (The AI Democracy Projects, Institute for Advanced Study, 2024), https://www.ias.edu/sites/default/files/Angwin-Nelson-Palta_SeekingReliableElectionInformationDontTrustAI_2-27-24.pdf.
- ⁵⁶ Messeri and Crockett, "Artificial Intelligence and Illusions of Understanding in Scientific Research."
- ⁵⁷ Harrison, "AI Accelerates Research Across Social Sciences"; and Christopher Barrie, Alexis Palmer, and Arthur Spirling, "Replication for Language Models: Problems, Principles, and Best Practice for Political Science," working paper, November 4, 2025, <https://www.cjbarrie.com/?p=1747f223485981578f86c6b41bf3e094&pm=c>.
- ⁵⁸ Nelson, *The Social Life of DNA*; and Benjamin, *Race After Technology*.

Philosophy of Autonomous Science: Ten Questions for the Coming Age of Artificial Scientists

Mario Krenn & Heather Champion

Artificial intelligence is beginning to make remarkable contributions in science, from protein design to materials discovery and experiment design. Yet we believe that moving from powerful tools to autonomous scientists is not just a technical challenge but also a deep philosophical and conceptual one. For that reason, we argue for a Philosophy of Autonomous Science (PAS): a program that translates core epistemic aims, including understanding, curiosity, surprise, interest, creativity, and novelty, into computable, nonanthropocentric objectives, while ensuring the safe and successful deployment of artificial scientists. Drawing on our own research experience, we show how insights from the philosophy of science can directly inform the design of AI systems for scientific discovery; outline ten inaugural questions that define the agenda of PAS; and invite philosophers, scientists, and AI researchers to collaborate in shaping the principles of the coming era of artificial scientists.

Artificial intelligence has started to contribute significantly to science. Let's consider biology, where Google DeepMind's AlphaFold program has solved one of the most significant challenges in structural biology – the protein prediction problem – leading to the 2024 Nobel Prize in Chemistry being awarded to DeepMind's John Jumper and Demis Hassabis (shared with biochemist David Baker for his work on computational protein design).¹ Similarly, in materials science, intelligent programs exploring the vast space of functional materials are beginning to discover molecules with exceptional properties.² In physics, AI algorithms have enabled, for the first time, the prediction of the sky location of neutron star mergers from gravitational wave signals before the mergers occur, opening the door to unprecedented multimessenger astronomy in the near future.³

Unquestionably, these are exciting achievements. However, if we aim to progress from AI being a useful scientific tool toward it being capable of autonomously conducting science, we must first recognize that this is not merely a technical challenge and it cannot be solved just by technical advances. Rather, it requires understanding how the best scientists perform science. For example, why and how are humans

creative? What constitutes scientific creativity? Why are humans curious, and what fundamental inner drives motivate scientists to deepen their understanding of the world? How do humans achieve understanding, and what precisely does scientific understanding entail? We believe that answering these questions and artificially recreating these traits is essential to the success of *autonomous science* (that is, science conducted by autonomous machine systems), even if not all of these human traits are replicable.

We strongly believe that the philosophy of science community can help answer these questions and thereby make significant contributions to the future of fully autonomous science. We propose a focused research program that studies how to translate traits of successful human researchers to autonomous machine scientists. The answers should be formulated in a nonanthropocentric way in the sense that they should be applicable to both biological and artificial scientists. We also encourage the study of the potential consequences of autonomous science, for example, on scientific understanding, one of the main aims of science.

To explain why we are so convinced that insights from philosophy could advance the automation of science, we offer our first personal anecdote. One main direction of our research (namely, that of Mario Krenn) is the automated discovery of new physics experiments.⁴ Our AI programs have discovered experimental setups that no human physicists could find (such as the generation of complex quantum entanglement or new quantum transformations, or the design of extremely sensitive gravitational wave detectors), and numerous experiments have been implemented in laboratories. Whenever a machine finds a solution to a question that humans cannot answer, that solution must contain new ideas and tricks that humans can learn from. The problem is that learning from and understanding computer-designed solutions can be extremely challenging. While it has recently been possible to do so – for example, with our 2017 discovery of a new way to generate photon entanglement – it usually takes weeks to understand the underlying physical processes.⁵ Around the year 2019, we started to think more deeply about how we can better gain new understanding with artificial intelligence – both for our particular questions and in general. We soon realized that because we could not consistently spell out what scientific understanding is, we couldn't even begin to answer the question. At this point, we discovered the work of Dutch philosopher Henk W. de Regt, who, in his 2017 book *Understanding Scientific Understanding* (which earned him the 2019 Lakatos Award for outstanding contribution to the philosophy of science), elucidated exactly what we were searching for: clear-cut and nonanthropocentric criteria to identify what scientific understanding is.⁶ De Regt's insight that a core principle of scientific understanding is the ability to apply new insights in different scenarios – that is, to generalize without full computation – helped us to develop an entirely new algorithm. We built the algorithm, Theseus, to enable humans to use the computer-designed solutions to quantum physics problems in different

contexts.⁷ We translated this goal into the application of a very abstract representation of quantum optics: based on graphs, together with highly efficient pruning and simplification techniques that produce solutions that conventionally can be understood immediately. Theseus allowed us to gain new insights into the principles of quantum entanglement generation and quantum interference. For example, by solving a question about the generation of heralded quantum entanglement for qubits, we immediately recognized the underlying structure that made the solution work and were able to generalize it to high-dimensional heralded entanglement generation that could be important in future quantum networks. We introduced Theseus in 2021 and still use the algorithm's generalizations for most quantum-optics design tasks today.⁸

This anecdote demonstrates that insights from the philosophy of science can directly contribute to the advancement and automation of physics. We are convinced that this was not a singular exception but that more insights from the philosophy of science could in the near future enable progress in AI for science and for autonomous science.

In this essay, we describe how the philosophy of science might change and expand its focus toward a new discipline, the *philosophy of autonomous science* (PAS), and how PAS could and should emerge, directly contributing to the design of artificial scientists to explore the universe, from its smallest to largest structures. While philosophers have discussed related issues, we believe their importance for autonomous science warrants far greater, sustained attention. For that reason, we propose ten crucial, initial research questions for this new field.

1 Scientific understanding. Imagine an oracle that correctly predicts the outcome of every particle physics experiment, the products of every chemical reaction, or the function of every protein. Such an oracle would revolutionize science and technology as we know them. However, as scientists, we would not be satisfied with the oracle alone. We desire more. We would want to comprehend how the oracle arrived at these predictions. This capability, known as *scientific understanding*, is frequently recognized as the essential goal of science.⁹ But what exactly does scientific understanding mean? How could artificial intelligence contribute to it? How could we recognize that an AI system possesses scientific understanding? And, finally, as AI systems become increasingly advanced, how can we continue to comprehend their superhuman solutions, concepts, and ideas?

Interestingly, notable philosophers in the early twentieth century developed comprehensive theories about scientific explanation while arguing that philosophers should disregard “understanding,” since these concepts are psychological and pragmatic rather than logical. Not all scientists agreed with this viewpoint. Physicists including Lord Kelvin, Erwin Schrödinger, and Richard Feynman developed their own models of scientific understanding.

In 2005, philosophers of science Henk W. de Regt and Dennis Dieks developed a new theory on what might constitute “scientific understanding,” foreshadowing de Regt’s work in *Understanding Scientific Understanding*.¹⁰ The de Regt-Dieks theory is contextual and pragmatic; it refers to observable properties rather than psychological states of mind. They regard previous concepts like mechanical models or visualizations as “tools for understanding” within a broader general framework. Their model, particularly in de Regt’s book, is strongly driven by the principle that a “satisfactory conception of scientific understanding should reflect the actual (contemporary and historical) practice of science.”

Their view is that “a phenomenon can be understood if there exists an intelligible theory T, such that scientists can qualitatively recognize characteristic consequences of T without performing exact calculations.” They explicitly define two interrelated criteria explaining what it means to understand a phenomenon and what constitutes an intelligible theory.

This theory is defined independently of humans; thus, it can – in principle – be applied to entities beyond humans, including machines. As such, it was a great inspiration for us to develop new algorithms and to classify scientific understanding through AI in general. However, one of the most powerful ideas in the definition – “without performing exact calculations” – makes a direct application to AI algorithms difficult. For that reason, we have transformed their definition into a criterion for teaching (see question eight), related to a Turing test. Unfortunately, this method cannot be scaled. Thus, large-scale exploration of AI-understanding cannot be tested. For that reason, one critical question is whether we can find one (or even more) nonanthropocentric and computable approximation of scientific understanding. And if so, what is its relation to de Regt’s criteria? Is it possible to maximize the new computable notion of scientific understanding?

2 Scientific motivations. What drives human researchers, such that they get up every day and work tirelessly to explore the world? At the core of this drive lie *epistemic emotions*, which are directly connected to knowledge, learning, and understanding. For example, curiosity motivates us to explore, scientific interest guides the direction of exploration, and surprise and wonder sustain engagement and attention when we encounter the unexpected or unknown, while transforming uncertainty into fascination that keeps our mind open to new possibilities and deeper understanding. (We will discuss curiosity, surprise, and interest in subsequent questions.)

What are other motivations for scientists, and how can those be recreated artificially? For example, the epistemic emotion of awe and wonder is intertwined with cultural practices.¹¹ What could constitute an AI “culture” that leads to related emergent phenomena? Another motivation comes from direct competition between scientists – often at the verge of an (expected) breakthrough. How could one recre-

ate true competitive behavior between artificial systems? Another motivation for some scientists comes from the awareness of our own mortality and the reality that each discovery is made against the backdrop of a finite life. Could the introduction of artificial finiteness in machines lead to more motivation to explore the world?

These questions (and the questions in the subsequent three sections) are closely related to emotions, albeit epistemic ones. We essentially ask how such emotions could be recreated artificially. However, all current implementations of emotions are biological. Can their functional roles be captured *in silico*, or are certain forms of biological embodiment indispensable for (epistemic) emotions?

3 Scientific curiosity. Scientists are curious about the world, and this curiosity shapes what they research. But what does *curiosity* actually mean? In AI research, the term “curiosity” is used for exploring complex environments with sparse rewards. Think about a Super Mario game: to finish a level – the goal of the game – a large number of actions must be performed, and it is not immediately clear how or whether a specific action brings the player closer to the goal. This is troublesome not only for human players but also for AI agents playing similar games. To overcome the problem of sparse rewards, researchers introduced a different form of reward that is *intrinsic* to the agent.¹² Artificial agents are developed such that they want to be able to predict the consequences of their own actions in an environment. A good way for them to learn to do so is by taking actions for which they cannot predict the consequences well. The agents seem to mimic a simple form of curiosity; thus, researchers call this intrinsic reward a “curiosity-based” reward. Surprisingly, this intrinsic curiosity-based reward leads to agents that can play many computer games without relying on the scores provided by the game itself.¹³

This raises a number of important questions: How close is this form of artificial curiosity to human curiosity? Can human curiosity be described as motivation to act when the consequences are difficult to predict, or do we have entirely different forms of curiosity? An interest in knowing the unknown seems universally relevant. But perhaps curiosity should be distinguished from richer emotional experiences such as *wonder*, in which the unknowing agent also experiences a sense of admiration (for a variety of possible aesthetics).

Helen De Cruz, a philosopher of science, argued that awe and wonder have a defamiliarizing effect, allowing us to escape existing patterns of thought.¹⁴ This makes them extremely useful for science; for example, in prompting revolutionary changes. Nonetheless, as some scientific discoveries have accidental triggers, both fleeting interests in novel features of an environment as well as deeper desires for elating new knowledge seem relevant. Can we find models of curiosity that range in degree and kind, closer to the complex forms experienced by humans but that we can measure in a variety of arbitrary entities, including humans, animals, and machines?

4 Scientific surprise. Surprising insights often open new ways of thinking about a system. They not only help individual scientists learn unexpected facts but, on a collective scale, can trigger unforeseen breakthroughs. It is empirically demonstrated that surprising connections lead to impactful scientific results.¹⁵ But what does *surprise* mean, how can we identify and quantify it, and how can we maximize the surprise generated by AI agents in science?

Two computable proxies for the intuitive notion of surprise are frequently used in computer science: *Bayesian surprise* and *Shannon's surprisal*. Bayesian surprise measures how much a new observation changes your beliefs; it quantifies the information gained when updating from prior to posterior. Technically, it is the Kullback-Leibler divergence from the posterior to the prior, so it is always nonnegative and is zero only when the observation leaves your beliefs unchanged. Unlike Shannon's surprisal, which measures how unlikely the new data are under your current model, Bayesian surprise measures the size of the belief update. Bayesian surprise has previously been used in an influential experiment on computational creativity, as we will discuss in question six.

Quite clearly, these two information-theoretical interpretations of surprise do not capture the epistemic emotion described by philosophers and psychologists. But there are other ways to approximate surprise. For example, an empirical study of over two hundred scientists found that surprise often carries an aesthetic component, a *feeling of beauty or elegance* in the unexpected coherence of a new insight.¹⁶ Could and should this aesthetic component be recreated to capture the notion of surprise? Could surprise naturally emerge as a consequence of the maximization of other qualities, or do we need *artificial* surprise to build truly powerful artificial scientists?

5 Scientific interest. Humans put much effort, money, and brainpower into things that interest them. In science, interest is one of the main drivers of research direction and is thereby responsible for scientific progress on both the individual and global scale. But what does *interest* mean? What interests humans, and in particular scientists? And how could we recreate scientific interest in an automated way and build machines that are intrinsically interested in exploring the universe?

In psychology, "interest" is often considered an epistemic emotion, along with surprise, confusion, and awe.¹⁷ A standard psychological way of describing interest combines two concepts. First, the artifact needs to have high novelty-complexity. But novelty is not enough; it also needs to have high comprehensibility. That is, humans need to have the impression that they could understand the underlying idea.

Many AI researchers are attempting to build systems capable of drawing from millions of scientific papers to help scientists create new and exciting research ideas. But are those ideas actually interesting? In a number of these projects, researchers

have evaluated how experienced scientists rank the interest of AI-generated ideas. In one of our own projects, called SciMuse, we asked more than one hundred research group leaders in the Max Planck Society to help us evaluate how interesting the AI-generated ideas are.¹⁸ While we now have more than four thousand human-evaluated ideas, we cannot yet explain what makes an idea or discovery interesting.

This leaves us with several critical questions again: How can comprehensibility be approximated computationally? Can we achieve true scientific interest by maximizing complexity and comprehensibility? What are other computable criteria that generate interest? Does interest emerge from other intrinsic motivations such as surprise or curiosity? Is scientific interest domain-dependent or are there universal, field-independent criteria? And how can a definition of scientific interest be informed by large-scale sociological experiments that involve many successful scientists?

6 Scientific creativity. Creativity is one of humanity's highest achievements; it is the driving force behind our most remarkable innovations. Although it is often celebrated in the arts – in paintings, poetry, or music – creativity is equally crucial in science and technology, where imaginative thinking paves the way for groundbreaking discoveries. But what is *creativity*? Does it require a special process like intentional reasoning, and, if so, can this be approximated with computational methods? How can we evaluate and potentially automate it in the scientific context?

Philosophers and scientists have long identified creativity using two crucial properties: novelty and usefulness. (Sometimes the closely related properties of originality and effectiveness are used, together with surprise, which has been advocated as a third criterion, to emphasize nontrivial forms of novelty.) This categorization is sometimes called the “standard definition of creativity.”¹⁹ In 2013, this abstract idea inspired a computational creativity experiment at IBM called Chef Watson. The goal was to develop an algorithm that could produce creative artifacts – in Chef Watson's case, creative culinary recipes and menus.²⁰ Through this experiment, the IBM team, in collaboration with chefs at the Institute of Culinary Education, worked to demonstrate how the two criteria of novelty and usefulness could be automatically evaluated and thus optimized. Toward these aims, the team collected a large dataset of ingredients, large datasets of international recipes of various dishes, and large datasets of experimental psychophysics data (human pleasantness ratings) and physicochemical features of flavor compounds. Novelty was estimated by a Bayesian surprise function – a Kullback-Leibler divergence that estimated how much a previously unseen dish changes the general recipe distribution. Usefulness was approximated by a score that estimated the pleasantness of the new dish computed from the large database of experimental pleasantness and flavors. The team then maximized combinations of the two

approximated concepts and indeed found numerous novel and pleasant recipes, as ranked by human experts (chefs).

This groundbreaking demonstration shows that complex ideas such as creativity can be approximated and maximized. Thus, the question of how we can approximate creativity in the natural science domain becomes crucial. Are there different forms of computable creativity, and what distinguishes artifacts that maximize one or the other approximation of creativity? Can different modes and qualities of creativity, such as those described by philosopher and cognitive scientist Margaret Boden – exploration of, association within, or transformation of a conceptual space – be computationally approximated?²¹ Additionally, could creativity emerge as a side effect of a more general optimization strategy, such as surprise maximization or uncertainty minimization?

7 Scientific novelty. As scientists, we strive for novel results. This sounds simple – but what is *novel*, and what is *novelty* in science? Sometimes, a tiny modification of a system has severe consequences. Sometimes, in hindsight, a discovery looks obvious, even if it was overlooked for a long time. This question of novelty is not only useful by itself, but as a key component of other crucial traits such as creativity or interest; we need to understand it well.

There are many metrics that capture some notions of scientific novelty. For example, in information theory, the number of bits that are required to encode a new artifact under the current model approximates novelty. In knowledge graphs or latent spaces of large neural networks, the distance of the new artifact to other known artifacts is used to approximate novelty. In AI-driven idea generation, novelty is sometimes captured by the nonexistence of links between concepts in knowledge graphs. These notions clearly only capture a small aspect of novelty in science. Recently, large language models (LLMs) have been used to automatically evaluate whether findings, ideas, or hypotheses are new – a process that clearly introduces new biases by prioritizing a particular concept of novelty.

In scientific exploration, researchers seek to discover new phenomena that are not specifically predicted by theory, emergent phenomena that are not directly visible from their discipline’s governing equations (for instance, superconductivity and the whole idea of quantum computing from Schrödinger’s equation), or new equations and models that describe experimental data. So what types of novelty exist and how do they indicate success? One of us (Heather Champion) recently proposed several types of “strong novelty” for science that demonstrate key learning outcomes, and showed how current AI algorithms are already achieving many of them by generating surprise, reducing utility blindness, and eliminating deep ignorance, while potentially changing scientific concepts.²²

In general terms, we ask: How can one identify novelty independent of the domain? And can general notions of scientific novelty be computable?

8 Machine-to-human pedagogy and AI teachers. In 2000, science fiction writer Ted Chiang wrote a prophetic one-page fiction story in *Nature* titled “Catching Crumbs from the Table.” In this story, Chiang describes the emergence of “metahumans”: genetically modified humans with enormous brain capacities. He portrays a world in which the last original human-authored research was submitted for publication twenty-five years prior, with humans unlikely ever to make any original scientific contributions again. In Chiang’s scenario, some human scientists “left the field altogether, but those who stayed shifted their attentions away from original research and toward hermeneutics: interpreting the scientific work of metahumans.”²³

We believe we are on the verge of entering such a world. We observe it in our own research into using intelligent algorithms to design new physics experiments. Traditionally, creative and experienced human researchers devise such experimental designs. However, given the enormous number of possible experimental setups, it is questionable whether humans have already discovered all designs with exceptional properties (our own research indicates they have not). As our human-machine team competes against the best human-created designs – crafted through careful, rational design decisions – we know that if machine solutions outperform human creations, those solutions must contain new ideas and concepts from which humans can learn. (The machine solutions can be directly verified by external tools that are not dependent on the algorithm that discovered them.) And indeed, in many cases, we have discovered new generalizable concepts and ideas in quantum physics by analyzing what the machines have discovered (we describe one such instance in the introduction to this essay).²⁴ However, this is not always the case. With physics researcher Yehonathan Drori and experimental physicist Rana Adhikari, we recently used AI algorithms to design new gravitational wave detectors, some of the most sensitive measurement devices ever conceived by humanity. These detectors have, so far, been designed by a collaboration of hundreds of researchers over the last twenty to thirty years. Our AI system has discovered more than a dozen new and at least theoretically more sensitive detector designs.²⁵ However, even after analyzing them for half a year, we were not able to understand the underlying big picture of the idea. One can compute the designs by hand and with software, but we failed to understand why they worked.

It seems inevitable that this dynamic will increasingly occur in physics and other sciences: Machines will invent solutions from which humans could learn. If we are fortunate (as in the first quantum optics example), we will grasp the underlying principles and thereby enhance our scientific understanding.²⁶ However, we anticipate that we will increasingly encounter scenarios similar to the gravitational wave detectors, where we can confirm but not fully comprehend the principles behind AI-discovered improvements. This raises the critical question: *What should we do?* Perhaps “we need not be intimidated by the accomplishments of metahuman

science,” as Chiang concludes in his prophetic text. Rather “we should always remember that the technologies that made metahumans possible were originally invented by humans, and they were no smarter than we.”

What else could be done to solve this gap in understanding AI-discovered solutions? One powerful option would be developing AI systems capable of explaining the underlying concepts and ideas to us. Such systems would go far beyond the current scope of interpretable AI or explainable AI (XAI). In XAI, the goal is for humans to understand the decision-making processes of advanced AI, which is crucial for AI safety and ethics.

However, it is unclear whether looking inside the “machine’s brain” is the most effective way for us to gain new understanding. Even if we can comprehend its internal workings, it remains uncertain whether humans will ever fully understand solutions in the same way as a system with superhuman intelligence. Our limited mental capacities in terms of memory and computational speed compel us to seek simplified models to understand phenomena. Conversely, an advanced system with access to different computational resources could develop scientific models beyond our capabilities.

Instead of examining the internal workings of powerful AI, we might look at how we learn from each other. When we want to learn what another human knows, we do not look into their neural circuits; rather, we communicate and learn from the teachings of other people. We might need to do the same for artificial systems: develop AI algorithms that can explain new physics models, concepts, and ideas at a level comprehensible to us. This capability closely aligns with the test for scientific understanding described in question one. In short, we should focus on building a Teacher AI (TAI) for science.

We can envision a TAI that either develops solutions or uses solutions from other AIs and explains them to humans. It could resemble a student-teacher dynamic, in which students ask questions for clarification and details, and the teacher provides insights. Likely, the TAI would need access to advanced scientific tools similar to those that humans use for visualization, simplification, or generalization. The TAI could generate exercises to test our knowledge and track our understanding and its progression during teaching.

Developing such a system also requires the TAI to have access to the same mathematical and physical tools used by human scientists. For example, in quantum mechanics, it would benefit from autonomous access to open-source programs like Qiskit (IBM’s quantum circuit simulator), NetKet (physicist Giuseppe Carleo’s toolbox for simulating many-body quantum systems), QuTiP (a quantum optics simulator developed by scientist Franco Nori’s research group at the RIKEN institute), and PyTheus (a quantum physics experiment discovery framework developed by our Krenn Research Group then-hosted at the Max Planck Institute for the Science of Light).²⁷ The TAI should not only provide verbal explanations but also fully

leverage the numerous “tools for understanding,” as described by de Regt.²⁸ For instance, regarding visualization, the TAI should determine how concepts and ideas can be visually represented through graphs, diagrams, or potentially videos and 3D animations. It should use unification by generating new examples that employ the same concepts and create practical exercises to help us grasp the application of these concepts. We have not seen practical effort in this direction, but in order to avoid Chiang’s dystopian future, such systems will be essential in the forthcoming years.

This potential raises numerous questions: How could machine-to-human pedagogy advance to ease human understanding of alien concepts? And what motivates human scientists to understand the universe when their role is degraded from active explorer to passive student?

9 Automating the philosophy of science. The philosophy of science has long shaped the rules of research, from Karl Popper’s theory of falsification and Imre Lakatos’s research programs on falsificationism to Paul Feyerabend’s pluralism and methodological anarchism.²⁹ If we reason about autonomous science, we must think about the underlying foundations and rules of science, and should consider the possibility of automating the philosophy of science itself. But what does that mean?

One potential way to automate some aspects of the philosophy of science would be to approximate “the rules of doing science” as policies that describe ways of generating hypotheses, developing overthrowing theories, or designing experiments. Through simulated worlds, these policies could be evaluated on the quality of the systems’ discoveries about the (simulated) worlds. The evaluation metrics could be formalized as objective functions to select high-performing policies – or, in the pluralistic spirit of Feyerabend, a diverse set of them – and thereby drive the discovery of new rules for exploring the world. Yet, as always, many questions arise here: What are good metrics for evaluating different policies, such as different rules of science? How can we guard against Goodhart’s law (policies that game the metric) and ensure that this autophilosophy preserves desirable traits such as methodological pluralism?³⁰ How can these policies be modeled and parametrized? And can large surveys of the beliefs of working philosophers and scientists help to model and emulate their decision processes?³¹

10 The role of humans in a future of autonomous science. Let us assume a world with successful artificial scientists that can explore the universe independently, without human input. What, then, is our role in such a world? Humans will remain the strategic leaders, setting the grand directions and determining how resources are allocated. These decisions will not be purely scientific; they will also be guided by societal goals, ethical considerations, and collective needs. Developing effective and transparent mechanisms for such

high-level decision-making will be one of the central challenges for the future governance of science.

At the same time, it will be essential to establish rigorous methods that allow humans to verify, interpret, and safely validate the discoveries proposed by artificial scientists. Is this possible if we cannot genuinely understand the AI-generated ideas? How can we maximize the benefits of autonomous scientific discovery while mitigating the risks of potentially catastrophic outcomes (for instance, the accidental discovery of an easily reproducible bioweapon or an algorithm capable of breaking modern encryption schemes)? Equally important will be ensuring that artificial scientists do not develop systematic blind spots: regions of the scientific search space left unexplored not because they are uninteresting but because of hidden architectural biases or misaligned objectives.

As the era of artificial scientists approaches, we must now begin to identify which human roles, values, and forms of oversight will remain indispensable in guiding, interpreting, and safeguarding the future of knowledge itself.

We have proposed ten questions that could – and should – form the foundation of a new field: the philosophy of autonomous science. The goal of this field should be to help develop and guide the emerging era of artificial scientists by translating fundamental epistemic traits and values (such as curiosity, creativity, motivation, surprise, interest, understanding, and novelty) into nonanthropocentric, computable objectives, and by establishing norms, governance frameworks, and control mechanisms for autonomous scientific systems.

If we build them with care, artificial scientists will extend our capabilities for curing diseases, solving mathematical mysteries, and exploring the cosmos at a speed beyond imagination.

ABOUT THE AUTHORS

Mario Krenn is Professor of Machine Learning in Science at the University of Tübingen, where he leads the Artificial Scientist Lab (previously based at the Max Planck Institute for the Science of Light). He has published in such journals as *Physical Review X*, *Machine Learning: Science and Technology*, *Nature Machine Intelligence*, and *Nature Photonics*.

Heather Champion is a Guest Researcher in the Ethics & Philosophy Lab, part of the Cluster of Excellence–Machine Learning for Science at the University of Tübingen. She is also a PhD candidate in Philosophy at the Rotman Institute of Philosophy and at Western University. She has published in such journals as *Synthese* and *Philosophy of Science*.

ENDNOTES

- ¹ John Jumper, Richard Evans, Alexander Pritzel, et al., “Highly Accurate Protein Structure Prediction with AlphaFold,” *Nature* 596 (7873) (2021): 583–589.
- ² Robert Pollice, Gabriel dos Passos Gomes, Matteo Aldeghi, et al., “Data-Driven Strategies for Accelerated Materials Design,” *Accounts of Chemical Research* 54 (4) (2021): 849–860.
- ³ Maximilian Dax, Stephen R. Green, Jonathan Gair, et al., “Real-Time Inference for Binary Neutron Star Mergers Using Machine Learning,” *Nature* 639 (8053) (2025): 49–53.
- ⁴ Mario Krenn, Mehul Malik, Robert Fickler, et al., “Automated Search for New Quantum Experiments,” *Physical Review Letters* 116 (9) (2016): 090405; and Mario Krenn, Manuel Erhard, and Anton Zeilinger, “Computer-Inspired Quantum Experiments,” *Nature Reviews Physics* 2 (11) (2020): 649–661.
- ⁵ Mario Krenn, Armin Hochrainer, Mayukh Lahiri, and Anton Zeilinger, “Entanglement by Path Identity,” *Physical Review Letters* 118 (2017): 080401.
- ⁶ Henk W. de Regt, *Understanding Scientific Understanding* (Oxford University Press, 2017); and The London School of Economics and Political Science, “Lakatos Award Lecture,” <https://www.lse.ac.uk/philosophy/events/lakatos-award-lecture> (accessed January 30, 2026).
- ⁷ Mario Krenn, Jakob S. Kottmann, Nora Tischler, and Alán Aspuru-Guzik, “Conceptual Understanding through Efficient Automated Design of Quantum Optical Experiments,” *Physical Review X* 11 (2021): 031044.
- ⁸ Carlos Ruiz-Gonzalez, Sören Arlt, Jan Petermann, et al., “Digital Discovery of 100 Diverse Quantum Experiments with PyTheus,” *Quantum* 7 (2023): 1204.
- ⁹ Henk W. de Regt and Dennis Dieks, “A Contextual Approach to Scientific Understanding,” *Synthese* 144 (1) (2005): 137–170; and Angela Potochnik, *Idealization and the Aims of Science* (University of Chicago Press, 2017).
- ¹⁰ De Regt and Dieks, “A Contextual Approach to Scientific Understanding”; and De Regt, *Understanding Scientific Understanding*.
- ¹¹ Helen De Cruz, *Wonderstruck: How Wonder and Awe Shape the Way We Think* (Princeton University Press, 2024).
- ¹² Jürgen Schmidhuber, “Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990–2010),” *IEEE Transactions on Autonomous Mental Development* 2 (3) (2010): 230–247; and Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell, “Curiosity-Driven Exploration by Self-Supervised Prediction,” in *ICML ’17: Proceedings of the 34th International Conference on Machine Learning – Volume 70*, ed. Doina Precup and Yee Whye Teh (Association for Computing Machinery, 2017), 2778–2787.
- ¹³ Yuri Burda, Harri Edwards, Deepak Pathak, et al., “Large-Scale Study of Curiosity-Driven Learning,” paper presented at the Seventh International Conference on Learning Representations (ICLR 2019), New Orleans, Louisiana, May 6–9, 2019.
- ¹⁴ Taylor Carman, “Wonderstruck: How Wonder and Awe Shape the Way We Think,” *Notre Dame Philosophical Reviews*, June 6, 2025; and De Cruz, *Wonderstruck*.
- ¹⁵ Feng Shi and James Evans, “Surprising Combinations of Research Contents and Contexts Are Related to Impact and Emerge with Scientific Outsiders from Distant Disciplines,” *Nature Communications* 14 (2023): 1641.

- ¹⁶ Milena Ivanova and Brandon Vaidyanathan, "Surprise in Science: A Qualitative Study," *Erkenntnis: An International Journal of Scientific Philosophy* 91 (1) (2026): 57–80.
- ¹⁷ Paul J. Silvia, "Interest—the Curious Emotion," *Current Directions in Psychological Science* 17 (1) (2008): 57–60.
- ¹⁸ Xuemei Gu and Mario Krenn, "Interesting Scientific Idea Generation Using Knowledge Graphs and LLMs: Evaluations with 100 Research Group Leaders," arXiv (2024), <https://doi.org/10.48550/arXiv.2405.17044>.
- ¹⁹ Mark A. Runco and Garrett J. Jaeger, "The Standard Definition of Creativity," *Creativity Research Journal* 24 (1) (2012): 92–96.
- ²⁰ Lav R. Varshney, Florian Pinel, Kush R. Varshney, et al., "A Big Data Approach to Computational Creativity," arXiv (2013), <https://doi.org/10.48550/arXiv.1311.1213>; and Lav R. Varshney, Florian Pinel, Kush R. Varshney, et al., "A Big Data Approach to Computational Creativity: The Curious Case of Chef Watson," *IBM Journal of Research and Development* 63 (1) (2019): 7.1–7.18, <https://doi.org/10.1147/JRD.2019.2893905>.
- ²¹ Margaret A. Boden, "Computer Models of Creativity," *AI Magazine* 30 (3) (2009): 3; and Margaret A. Boden, *The Creative Mind: Myths and Mechanisms*, 2nd ed. (Routledge, 2004). As Boden explains in her introduction to *The Creative Mind*, reprinted by *Interalia Magazine* in 2016: "'What is going on' isn't magic—and it's different in each type of case. For creativity can happen in three main ways, which correspond to the three sorts of surprise. The first involves making unfamiliar combinations of familiar ideas. Examples include poetic imagery, collage in painting or textile art, and analogies. These new combinations can be generated either deliberately or, often, unconsciously. . . . The other two types of creativity are interestingly different from the first. They involve the *exploration*, and in the most surprising cases the *transformation*, of conceptual spaces in people's minds. Conceptual spaces are structured styles of thought. They're normally picked up from one's own culture or peer-group, but are occasionally borrowed from other cultures." See Margaret A. Boden, "Creativity in a Nutshell," *Interalia Magazine*, <https://www.interaliamag.org/articles/margaret-boden-creativity-in-a-nutshell> (accessed February 5, 2026).
- ²² Heather Champion, "Strong Novelty Regained: High-Impact Outcomes of Machine Learning for Science," *Synthese* 206 (3) (2025): 1–23.
- ²³ Ted Chiang, "Catching Crumbs from the Table," *Nature* 405 (6786) (2000): 517, <https://doi.org/10.1038/35014679>.
- ²⁴ Krenn, Hochrainer, Lahiri, and Zeilinger, "Entanglement by Path Identity."
- ²⁵ Mario Krenn, Yehonathan Drori, and Rana X. Adhikari, "Digital Discovery of Interferometric Gravitational Wave Detectors," *Physical Review X* 15 (2) (2025): 021012.
- ²⁶ Krenn, Malik, Fickler, et al., "Automated Search for New Quantum Experiments."
- ²⁷ IBM Quantum Computing, "Qiskit," <https://www.ibm.com/quantum/qiskit> (accessed February 6, 2026); Giuseppe Carleo, Kenny Choo, Damian Hofmann, et al., "NetKet: A Machine Learning Toolkit for Many-Body Quantum Systems," *SoftwareX* 10 (2019): 100311, <https://doi.org/10.1016/j.softx.2019.100311>; J. Robert Johansson, Paul D. Nation, and Franco Nori, "QuTiP: An Open-Source Python Framework for the Dynamics of Open Quantum Systems," *Computer Physics Communications* 183 (8) (2012): 1760–1772, <https://doi.org/10.1016/j.cpc.2012.02.021>; and Ruiz-Gonzalez, Arlt, Petermann, et al., "Digital Discovery of 100 Diverse Quantum Experiments with PyTheus."
- ²⁸ De Regt, *Understanding Scientific Understanding*.

- ²⁹ Karl Popper, *The Logic of Scientific Discovery* (Hutchinson & Co, 1959); Imre Lakatos, “Falsification and the Methodology of Scientific Research Programmes,” in *Criticism and the Growth of Knowledge: Proceedings of the International Colloquium in the Philosophy of Science, London, 1965*, ed. Imre Lakatos and Alan Musgrave (Cambridge University Press, 1970); and Paul Feyerabend, *Against Method: Outline of an Anarchistic Theory of Knowledge*, 4th ed. (Verso, 2010).
- ³⁰ Eric Oberheim, “Rediscovering Einstein’s Legacy: How Einstein Anticipates Kuhn and Feyerabend on the Nature of Science,” *Studies in History and Philosophy of Science Part A* 57 (2016): 17–26; and Charles Goodhart, “Problems of Monetary Management: The U.K. Experience,” in *Papers in Monetary Economics, 1975* (Reserve Bank of Australia, 1975).
- ³¹ David Bourget and David J. Chalmers, “What Do Philosophers Believe?” *Philosophical Studies* 170 (3) (2014): 465–500.

Geometry-Informed AI for Scientific Discovery

Melanie Weber

Artificial intelligence offers tremendous opportunities to accelerate scientific discovery. Already, AI models have driven major breakthroughs in areas such as protein folding and weather forecasting. Going forward, AI may hold the keys to the development of treatments for presently incurable diseases, to the design of innovative new materials, and to resolving decades-long open problems in mathematics. Currently, the model landscape is dominated by transformers; however, their extreme data and compute requirements and limited interpretability present fundamental challenges for the road ahead. Progress toward the next generation of AI for science will require a shift toward smaller, *structured* models that are efficient, interpretable, and capable of acting as genuine scientific collaborators: guiding experiments, revealing hidden patterns in data, and helping to chart new scientific frontiers.

One possible avenue for developing such next-generation models, which I am pursuing in my research, is *geometry-informed* AI.¹ This is motivated by the observation that encoding data geometry as inductive bias into models can mitigate high resource demands by reducing the model's data and computing needs, as demonstrated by recent work from my group and others.² To illustrate this, consider the example of image classification, in which labels are typically assigned based on the objects shown in the image, irrespective of their location. One can encode such structure into a model by only considering model architectures whose outputs are agnostic to shifts of objects within the image. Convolutional neural networks, which achieved a major breakthrough in image classification in the 2010s, are an early example of a model architecture that has this property.³ More broadly, models can encode a variety of geometric structures, such as symmetries arising from fundamental laws of physics or low-dimensional structure reflecting inherent correlations in high-dimensional data.

Why does this matter for the sciences? Take state-of-the-art weather prediction models, which require training on millions of examples to provide accurate forecasts. By encoding known geometric structure, we can prevent models from expending resources on implausible scenarios. This is accomplished by constrain-

ing model outputs to remain consistent with established domain knowledge, such as physical laws that enforce specific symmetries.

Geometric models are already driving progress across scientific domains: Equivariant models accelerate materials discovery and physics-informed architectures achieve high-resolution weather forecasting.⁴ In my own interdisciplinary work, we have developed geometric approaches for representing large-scale single-cell data in a way that makes it easier for researchers to trace how cells mature and branch into different types in developmental processes.⁵ For astronomy, we developed a method that computes interpretable representations of galaxies that preserve information about their surrounding cosmic environment, uncovering patterns tied to galaxy mass and star formation.⁶ In structural biology, we introduced a geometry-aware framework for heterogeneous cryo-EM reconstruction that predicts atomic backbone conformations, allowing researchers to recover protein structural variability from single-particle images using geometric priors.⁷

State-of-the-art geometric models can already rival transformer-based architectures in specialized tasks, yet they represent only an early glimpse of what geometry-informed AI could enable. To move beyond specialized solutions and toward more versatile AI assistants, both mathematics and engineering advances are needed. Recent theoretical work by my group and others has begun to clarify when and why geometric priors yield provable advantages, but deeper mathematical foundations, and their incorporation into broadly applicable foundation models, will be key to accelerating progress in geometry-informed AI.⁸ Equally important is the development of architectures that encode richer and more heterogeneous geometric structure, along with the software and hardware infrastructure needed to scale them efficiently.⁹ Together, these efforts could enable geometric models capable of driving major advances in AI-assisted scientific discovery.

While I have emphasized that mathematical insights can advance artificial intelligence, the converse is increasingly true as well. AI-assisted reasoning tools are being introduced into mathematical research, with early successes showing that they are particularly effective for “needle in a haystack” problems, where a correct construction, counterexample, or identity exists but is extremely difficult to find with classical approaches.¹⁰ Since mathematics is fundamentally concerned with identifying and exploiting structure, I expect that structured models, such as geometry-informed AI, will play an important role going forward. Such approaches could mitigate some of the growing pains of current methods: They rely heavily on large language models that require time-consuming human formalization of mathematical data as text or draw on unreliable informal sources. Their outputs are also difficult for humans to interpret, limiting the extraction of proof strategies and mathematical intuition. Encoding mathematical principles directly into models could enable them to reason in ways that more closely align with human intuition,

strengthening human-AI collaboration in research mathematics. In my own research, we have already seen early successes with geometric models in combinatorics.¹¹

Looking ahead, I envision a scientific ecosystem in which AI is not merely a tool but a collaborator. While current models primarily assist with data analysis, future models could help researchers interrogate existing literature, act as sounding boards for hypothesis generation, and collaborate in the conception of new methodologies. Such human-AI collaborations have the potential to accelerate progress toward answering the great scientific questions of our time, and those yet to emerge.

ABOUT THE AUTHOR

Melanie Weber is an Assistant Professor of Applied Mathematics and of Computer Science at Harvard University, where she leads the Geometric Machine Learning Group. She is also a 2025 Early Career Fellow with the Schmidt Sciences AI2050 Project and a 2024 Research Fellow in Mathematics with the Alfred P. Sloan Foundation. Her research studies geometric structure in data and models and how to leverage such information for the design of new, efficient machine-learning methods with provable guarantees.

ENDNOTES

- ¹ Melanie Weber, “Geometric Machine Learning,” *AI Magazine*, January 10, 2025, <https://doi.org/10.1002/aaai.12210>.
- ² Alberto Bietti, Luca Venturi, and Joan Bruna, “On the Sample Complexity of Learning Under Invariance and Geometric Stability,” in *NeurIPS ’21: Proceedings of the 35th International Conference on Neural Information Processing Systems*, ed. Marc’Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, et al. (Curran Associates, Inc., 2021); Song Mei, Theodor Misiakiewicz, and Andrea Montanari, “Learning with Invariances in Random Features and Kernel Models,” *Proceedings of Machine Learning Research* 134 (2021); Bobak Kiani, Thien Le, Hannah Lawrence, et al., “On the Hardness of Learning Under Symmetries,” paper presented at the Twelfth International Conference on Learning Representations, Vienna, Austria, May 7, 2024, <https://openreview.net/forum?id=ARPrutzAnQ>; Bobak T. Kiani, Jason Wang, and Melanie Weber, “On the Hardness of Learning Neural Networks Under the Manifold Hypothesis,” in *NeurIPS ’24: Proceedings of the 38th International Conference on Neural Information Processing Systems*, ed. Amir Globerson, Lester Mackey, Danielle Belgrave, et al. (Curran Associates, Inc., 2024); and Johann Brehmer, Sönke Behrends, Pim De Haan, and Taco Cohen, “Does Equivariance Matter at Scale?” *Transactions on Machine Learning Research* (2025), <https://openreview.net/pdf?id=wilNute8Tn>.
- ³ Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, “Gradient-Based Learning Applied to Document Recognition,” *Proceedings of the IEEE* 86 (11) (1998).
- ⁴ Simon Batzner, Albert Musaelian, Lixin Sun, et al., “E(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials,” *Nature Communications* 13

- (2022); and Jaideep Pathak, Shashank Subramanian, Peter Harrington, et al., “FourCast-Net: A Global Data-Driven High-Resolution Weather Model Using Adaptive Fourier Neural Operators,” arXiv (2022), <https://doi.org/10.48550/arXiv.2202.11214>.
- ⁵ Nithya Bhasker, Hattie Chung, Louis Boucherie, et al., “Uncovering Developmental Lineages from Single-Cell Data with Contrastive Poincaré Maps,” bioRxiv (2025), <https://doi.org/10.1101/2025.08.22.671789>.
- ⁶ Ana Sofia Uzoy, Claire Lamman, and Melanie Weber, “Manifold Learning for Cosmic Structures,” paper presented at the Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS 2025), San Diego, California, December 1, 2025, <https://neurips.cc/virtual/2025/loc/san-diego/122903>.
- ⁷ Jonathan Krook, Axel Janson, Joakim Andén, et al., “Protein Graph Neural Networks for Heterogeneous Cryo-EM Reconstruction,” paper presented at the IEEE International Conference on Image Processing, Tampere, Finland, September 13–17, 2026, <https://2026.ieeeicip.org>.
- ⁸ Bietti, Venturi, and Bruna, “On the Sample Complexity of Learning Under Invariance and Geometric Stability”; Mei, Misiakiewicz, and Montanari, “Learning with Invariances in Random Features and Kernel Models”; Kiani, Le, Lawrence, et al., “On the Hardness of Learning Under Symmetries”; and Kiani, Wang, and Weber, “On the Hardness of Learning Neural Networks Under the Manifold Hypothesis.”
- ⁹ Neil He, Jiahong Liu, Buze Zhang, et al., “Position: Beyond Euclidean – Foundation Models Should Embrace Non-Euclidean Geometries,” paper presented at Learning on Graphs Conference 2025, Arizona State University, December 10, 2025, <https://openreview.net/forum?id=WoK4090lln>.
- ¹⁰ François Charton, Jordan S. Ellenberg, Adam Zsolt Wagner, and Geordie Williamson, “Patternboost: Constructions in Mathematics with a Little Help from AI,” arXiv (2024), <https://doi.org/10.48550/arXiv.2411.00566>; Alex Davies, Petar Veličković, Lars Buesing, et al., “Advancing Mathematics by Guiding Human Intuition with AI,” *Nature* 600 (7887) (2021); and Alexander Novikov, Ngán Vű, Marvin Eisenberger, et al., “AlphaEvolve: A Coding Agent for Scientific and Algorithmic Discovery,” arXiv (2025), <https://doi.org/10.48550/arXiv.2506.13131>.
- ¹¹ Knut Vanderbush and Melanie Weber, “Neural Algorithmic Reasoning for Approximate k -Coloring with Recursive Warm Starts,” arXiv (2026), <https://arxiv.org/html/2601.05137v1>.

Are Current AI Systems Unlocking Knowledge Discovery in Genomics?

Antonio Orvieto

The last few years of building large foundation models in language and vision domains have revealed how steady progress can be achieved by scaling resources while keeping the model architecture (a softmax attention-based Transformer) fixed.¹ This strategy has been readily applied in science, where large-scale open-source foundation models have been released, and scaling laws indicate that pretraining (or non-task-specific) performance improves with increased compute and data availability.²

Despite the impressive versatility and zero- and few-shot performance of such AI-for-science models, recent studies have questioned whether they are genuine foundational architectures, and whether they truly let scientists process data with unprecedented accuracy compared to pre-2023 machine learning. While scaling can reveal pleasant surprises for the future, we now have evidence that simple supervised machine-learning approaches from the last decade – while less versatile and requiring sufficient training data – can still offer a performance edge over large foundation models at a fraction of the computational resources needed. For instance, computational biologist Ziqi Tang and colleagues evaluated the representational power of pretrained genomic foundation models for major functional genomics prediction tasks spanning DNA and RNA regulation, and found that probing their representations can offer no substantial advantage over machine-learning approaches from the last decade.³

This is, in a way, to be expected: the most profound aspect of modern AI systems is their versatility. Yet can progress in biological data processing come solely from enhanced versatility? As data availability increases, it is crucial not to follow text-driven AI insights uncritically, but to dedicate our efforts to rethinking and understanding how best to design and (pre)train specialized models with unprecedented predictive power for scientific tasks. A likely path forward, exemplified by recent progress in stateful memory models (such as Mamba and the xLSTM), is to shift away from the needs of pure language modeling toward neural networks inspired by human memory and smooth structures in natural data.⁴ It is helpful to remember that the transformer architecture was initially designed for the text

domain and in-context retrieval, and was later applied to other domains due to its ease of multimodal adaptation and the technical needs of our computational infrastructures. In light of recent literature showing concerning limitations of this neural network (such as state tracking, which relies on dialogue to create records and context for conversations over time), perhaps it is now time we abandon architectures defined by the needs of text and instead design truly domain-specific foundation models, beginning with the unique characteristics of scientific data.⁵

The DNA structure offers an excellent motivation for this paradigm shift: citing historian of biology Nathaniel Comfort, “if a genome is text, it is badly edited. Most DNA is gibberish, full of stutters, snippets of doggerel from other species.” DNA also offers a compelling example of a nontext domain in which data availability is rapidly increasing.⁶ The cost of sequencing is dropping sharply, and as genomic data volumes grow, AI gains enormous potential to reveal new knowledge. Yet we still lack understanding of the hidden logic in 98 percent of our genome that does not code for proteins – the noncoding regions whose patterns describe vital yet mostly unknown regulatory functions. Unveiling novel structures in these long sequences could unlock new treatments for cancer, autoimmune disorders, and neurological conditions.

In my theory-focused lab at the Max Planck Institute for Intelligent Systems, part of the newly established ELLIS Institute in Tübingen, Germany, we develop new architectures and training techniques for processing and pattern matching in extremely long sequences of symbols, such as those in our DNA.⁷ Specifically, we are developing new efficient models with adaptive hierarchical computation that can increase the reasoning budget based on data, ensuring mathematical guarantees of enhanced expressivity toward Turing-complete reasoning. Our latest Fixed-Point RNN model can surpass both transformers and long short-term memory networks (LSTMs) on a wide range of reasoning benchmarks, including state tracking, with no need for chain-of-thought reasoning or language modeling pretraining.⁸

Yet providing a good fit for the biological domain is not as straightforward as expected: benchmark results show only minor improvements, and we lack solid intuitions about the reasoning archetypes (memorization, recall, tracking) that could unlock understanding of long nucleotide sequences. Despite our clear direction toward improving long-sequence DNA processing machines – with enhanced expressivity through recurrent depth, hierarchical reasoning, and dynamical context pruning – the path forward can only be successful through close collaboration between AI experts and biologists to design robust, hard benchmarks that assess progress in a quantifiable way and direct researchers worldwide toward challenging problems.

So are current AI systems unlocking knowledge discovery in genomics? Perhaps not yet; but they are teaching us how to ask the right questions, illuminating

the path toward systems that eventually will. As artificial intelligence advances and data costs fall, we must approach new genomic technologies with an open mind: Progress will depend not on scaling alone but on our willingness to question assumptions. Rather than following insights from other domains, we should build a scientific foundation. Across my lab and many others, AI in genomics represents both a technical and philosophical opportunity – one that can advance drug discovery, benefit society, and push forward the theory of application-aware neural network design.

ABOUT THE AUTHOR

Antonio Orvieto is a Hector Endowed Principal Investigator at the ELLIS Institute Tübingen and an Independent Group Leader at the Max Planck Institute for Intelligent Systems, where he leads the Deep Models and Optimization group. He is also a Lecturer at the University of Tübingen, faculty for PhD programs in Tübingen and Zurich, and an AI2050 Early Career Fellow. He is an Action Editor for *Transactions on Machine Learning Research* and an Area Chair for NeurIPS and ICML. His papers have been accepted and published at such major AI conferences as NeurIPS, ICLR, ICML, and AISTATS.

ENDNOTES

- ¹ Jared Kaplan, Sam McCandlish, Tom Henighan, et al., “Scaling Laws for Neural Language Models,” arXiv (2020), <https://doi.org/10.48550/arXiv.2001.08361>.
- ² Žiga Avsec, Natasha Latysheva, Jun Cheng, et al., “Advancing Regulatory Variant Effect Prediction with AlphaGenome,” *Nature* 649 (8099) (2026): 1206–1218, <https://doi.org/10.1038/s41586-025-10014-0>; and Garyk Brixi, Matthew G. Durrant, Jerome Ku, et al., “Genome Modelling and Design Across All Domains of Life with Evo2,” *Nature* 652 (8112) (2026): 1349–1361, <https://doi.org/10.1038/s41586-026-10176-5>.
- ³ Ziqi Tang, Nirali Somia, Yiyang Yu, and Peter K. Koo, “Evaluating the Representational Power of Pre-Trained DNA Language Models for Regulatory Genomics,” *Genome Biology* 26 (1) (2025).
- ⁴ Albert Gu and Tri Dao, “Mamba: Linear-Time Sequence Modeling with Selective State Spaces,” paper presented at the First Conference on Language Modeling (COLM 2024), Philadelphia, Pennsylvania, October 7, 2024, <https://openreview.net/pdf?id=tEYskw1VY2>; and Maximilian Beck, Korbinian Pöppel, Markus Spanring, et al., “xLSTM: Extended Long Short-Term Memory,” in *NIPS ’24: Proceedings of the 38th International Conference on Neural Information Processing Systems*, ed. Amir Globerson, Lester Mackey, Danielle Belgrave, et al. (Curran Associates, Inc., 2024).
- ⁵ William Merrill, Jackson Petty, and Ashish Sabharwal, “The Illusion of State in State-Space Models,” in *ICML ’24: Proceedings of the 41st International Conference on Machine Learning*, ed.

Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, et al. (Journal of Machine Learning Research, 2024).

- ⁶ Nathaniel Comfort, “Genetics: We Are the 98%,” *Nature* 520 (7549) (2015): 615–616.
- ⁷ Antonio Orvieto, Samuel L. Smith, Albert Gu, et al., “Resurrecting Recurrent Neural Networks for Long Sequences,” in *ICML '23: Proceedings of the 40th International Conference on Machine Learning*, ed. Andreas Krause, Emma Brunskill, Kyunghyun Cho, et al. (Journal of Machine Learning Research, 2023).
- ⁸ Sajad Movahedi, Felix Sarnthein, Nicola Muca Cirone, and Antonio Orvieto, “Fixed-Point RNNs: Interpolating from Diagonal to Dense,” paper presented at the Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS 2025), San Diego, California, December 1, 2025, <https://openreview.net/pdf?id=KT8y9pFgJE>.

AI & the Discovery of Molecules through Autonomous Laboratories

Connor Coley

We need new molecules. Medicines for diseases lacking treatments and for future pandemics, catalysts to decarbonize chemical industries and enable plastic recycling, stabilizers for battery longevity and energy storage, and agricultural agents to protect crops from blights and invasive species. Finding such molecules is a search. We *design* candidates based on what we know, then *evaluate* them in the lab, then learn and repeat. Artificial intelligence has accelerated design, but the bottleneck to practical impact is not a lack of ideas. It is the hard, tactile work of making and testing new physical matter.

As much as we hope to simulate the behavior of molecules in the digital world, every proposal eventually collides with the physical world. And in the physical world, things rarely go according to plan. Many molecules, including those with exquisitely complex structures found in nature, can take months or years to synthesize. When finally made, top-scoring molecules designed by AI often underperform.¹ Even as our ability to design, score, and rank structures improve, physical experiments remain the final arbiter. The practical question is not merely whether AI can propose molecules but whether these proposals can be efficiently translated into reality.

This “execution gap” is the central challenge of autonomous chemical laboratories – facilities that seamlessly close the loop between hypothesis and experiment – and has shaped how I view opportunities in AI for science for the past decade. What good is it to design molecules we cannot practically test? Chemistry offers a vivid case study, but the challenge echoes across biology and other scientific fields. The community has taken notice of this question and is slowly shifting focus from *design* to *execution*. Excitement has grown around the concept of autonomous laboratories with enough optimism to fuel hundred-million-dollar seed rounds.

The impulse behind these efforts is older than AI. Autonomous laboratories are part of a long tradition of scientists searching for new scientific tools that improve our capacity to measure, reason, and design.

My path to the use of AI in chemistry began in chemical engineering. Chemical engineering is often associated with the petroleum industry but is fundamentally about using quantitative reasoning to understand and

optimize physical processes from atomic to global scales. Optimizing chemical transformations for the pharmaceutical industry using design of experiments and statistical techniques progressed into cheminformatics and machine learning.

During my graduate work, I split my days between code and hardware by writing algorithms for planning synthesis recipes of new molecules while constructing physical platforms that could, in principle, execute those plans. The culmination of one large collaborative effort was a demonstration of manufacturing drug molecules through AI-designed reactions carried out by a robotic platform. It worked, but admittedly only within narrow boundaries. What looks simple on a whiteboard becomes far less so at the lab bench. The gap between the level of detail in today's AI-generated recipes and what is needed to make them real recalls the classic children's exercise of writing instructions for a robot to make a peanut butter and jelly sandwich.

Autonomous laboratories are, in essence, *embodied AI* systems that link design to evaluation. Our physical laboratories are the hands (and eyes) of our AI designers. Our best autonomous laboratories are currently fenced-in explorers that can move quickly and reliably inside a well-defined search space, where the scientific hypotheses they are allowed to generate are limited to what they are equipped to test. Similar approaches have already borne fruit in fully digital domains where AI systems learn within clearly defined boundaries and evaluate their own progress through simulation, as seen in efforts from game play to Google's foray into "empirical software."²

The current and perennial challenge is to expand the playing field for AI design tools. The new molecules we need to meet pressing societal challenges are unlikely to be within the capabilities of today's push-button systems. Progress will depend on connecting three fronts: better planning of detailed synthetic routes and experimental procedures, richer perception through sensing and real-time data interpretation, and orchestration systems that translate protocols into instrument commands, coordinate queues, and adapt to failures. The physical world adds complications – sticky powders, viscous liquids, sensitive intermediates – that will require advances in robotic dexterity and reliability. Even successful reactions can produce ambiguous results and, as the A-Lab study shows, understanding what those results mean is a challenge of its own.³ The success of autonomous laboratories cannot be measured by the number of experiments performed or terabytes of information acquired per day. It must be measured by how effectively it accelerates learning through tight design cycles, faster loops, and fewer dead ends. These problems and pursuits lack the clean success metrics and benchmarking datasets that propelled progress in computer vision and protein structure prediction, yet they are essential to progress.⁴

The long arc of this field points toward a new style of scientific infrastructure: robust, broadly capable autonomous laboratories that link physical and digital worlds. And suppose we close the execution gap – What then? What happens when

laboratories can learn from every experiment, plan new ones on their own, and operate with the reliability of the best human scientists? In chemistry and medicine, such systems could map unexplored regions of chemical space, uncover new reaction pathways, and generate hypotheses that no individual researcher could have enumerated. They might shorten the distance between molecular insight and therapeutic impact, allowing us to test ideas that would otherwise be dismissed as too slow, too expensive, or too uncertain to pursue. Requesting a physical experiment might not feel so different from requesting a long-running simulation. More capital intensive and still with environmental impact considerations, yes, but no less convenient. Human scientists would be freed to focus on defining the goals, priors, and ethics of scientific inquiry.

Investing in this future is essential if we are to discover new medicines and materials at the pace the world requires. Opportunities beyond efficiency lie in using these tools to ask better questions, not merely to find faster answers.

ABOUT THE AUTHOR

Connor Coley is the Class of 1957 Career Development Professor and Associate Professor in the Department of Chemical Engineering and the Department of Electrical Engineering and Computer Science at MIT. He leads a research group that integrates chemistry and artificial intelligence to develop data-driven models that describe how molecules behave, interact, and react for the purposes of molecular design, synthesis, and analysis. He has published in such journals as *Nature*, *Science*, and *Nature Catalysis*.

ENDNOTES

- ¹ We will set aside the additional complexities that arise when progressing from the “in vitro” lab to, for example, clinical trials.
- ² Lizzie Dorfman and Michael Brenner, “Accelerating Scientific Discovery with AI-Powered Empirical Software,” Google Research (blog), September 9, 2025, <https://research.google/blog/accelerating-scientific-discovery-with-ai-powered-empirical-software>.
- ³ Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei, et al., “An Autonomous Laboratory for the Accelerated Synthesis of Inorganic Materials,” *Nature* 624 (7990) (2023): 86–91; and Josh Leeman, Yuhao Liu, Joseph Stiles, et al., “Challenges in High-Throughput Inorganic Materials Prediction and Autonomous Synthesis,” *PRX Energy* 3 (2024): 011002.
- ⁴ See Jia Deng, Wei Dong, Richard Socher, et al., “ImageNet: A Large-Scale Hierarchical Image Database,” 2009 *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society, 2009); and Helen M. Berman, John Westbrook, Zukang Feng, et al., “The Protein Data Bank,” *Nucleic Acids Research* 28 (1) (2000): 235–242.

Scaling Physics Intelligence for the Earth's Subsurface

Gege Wen

The progress of AI in the last decade has come with an enormous cost: energy. Modern AI development resembles a double Ouroboros, the Greek mythological symbol of two snakes biting each other's tails. To scale AI, we must reinvent the energy system, and to reinvent the energy system, we must harness the power of AI.

AI is a greedy consumer of power. Training a single large foundation model can consume as much electricity as thousands of U.S. households use in a year. What's more, AI doesn't just require "energy" broadly; it needs firm, stable, responsive power capable of meeting sudden compute spikes and massive parallel processing. Unfortunately, intermittent renewables like solar and wind alone cannot deliver that. The scalable and economically viable solution lies in Earth's subsurface: a mixture of natural gas-fired power plants with carbon sequestration, large-scale energy storage for renewables, and geothermal power.

But these subsurface energy systems are among the most complex and uncertain environments humans have ever tried to engineer. To design and operate them effectively, we need predictive tools that can understand how physics behaves in the subsurface, including how fluids travel within porous media, where carbon dioxide injection will cause high pressure, and whether heat can be produced sustainably in enhanced geothermal systems. All of these complex physics are occurring in the opaque, heterogeneous, and deep subsurface. That's a level of complexity that can't yet be properly captured with today's technology.

The digital toolkit that supports subsurface engineering today was built during the oil and gas boom of the 1990s and early 2000s. Having worked with these legacy systems early in my career, I've experienced firsthand their fundamental limitations. As an intern on a typical engineering team – comprising one manager, one geologist, one reservoir engineer, and one operation engineer on site – my main weekly task was to print out the production histories using Excel spreadsheets. On Mondays, we gathered in a small office, staring at columns of numbers and a large, printed map, debating where to send the drilling rig next. In this system, drilling

an exploratory well had a success rate of about 50 percent. This high failure rate shaped an ecosystem of tools designed around caution, not accuracy. Simulations could take weeks. Models were calibrated manually. Uncertainty wasn't something to eliminate; it was simply priced into the business.

These workflows were never meant to support a world in which we must quickly evaluate hundreds of potential sites or simulate multiphase carbon dioxide behavior across an entire sedimentary basin in days. Nor were they meant to provide answers to AI engineers asking whether their next data center can run on stable carbon-free power 24/7.

And so we arrive at the Ouroboros: AI needs subsurface energy to grow, and subsurface energy needs AI to be harnessed. To build the infrastructure that powers AI, we need to fundamentally rethink the way we model subsurface physical systems.

It amazes me how much harder it is to learn to simulate the physical world than to generate text. Language, for all its richness, is fundamentally one-dimensional. Words flow linearly; their meanings are flexible. There are many ways to say the same thing. In other words, natural language is low entropy and high redundancy.

The physical world is the opposite. Every molecule, every parcel of fluid, every pocket of pressure is governed by strict constraints across four dimensions: three in space, one in time. Every point in the field must simultaneously satisfy conservation laws and boundary conditions. This is a level of accuracy that was not required previously in generative language and image models.

This is why simulating physics with AI is so challenging – and so exciting. If we can train AI systems to capture the rules of the real world, not just the patterns of language, we would unlock an entirely new paradigm. By integrating physics into today's AI, we can explore thousands of geological configurations and storage scenarios and iterate quickly on the design of future energy systems, all without waiting days for an oil and gas legacy numerical solver to finish.

My team explores such models: subsurface physics intelligence that predicts how gases and fluids flow through porous rock using synthetic simulations rooted in physical laws and real-world experimental datasets. These models can forecast the long-term performance of geothermal wells and identify optimal well locations for carbon dioxide injectors, drastically shortening the engineering process for designing such projects.

One success story is the web application ccsnet.ai, which we built to make AI for subsurface carbon dioxide storage broadly accessible. The platform hosts AI models pretrained with large numerical simulation datasets, covering wide ranges of initial and boundary conditions for subsurface physical systems. Once trained, each prediction takes a fraction of a second. By democratizing access to fast and high-quality reservoir simulation, the tool enables groups that previously lacked these capabilities – including small operators, national labs, and even individuals inter-

ested in local projects – to independently evaluate storage sites. Today, the web app serves about two hundred users per month from every continent except Antarctica.

Of course, such work still needs to be scaled up to support industrial use. But it already offers a glimpse of what the energy industry could look like in the next five to ten years. Early in my career, a typical day meant going into the office, turning on my computer, setting up a numerical simulation case, and waiting – sometimes for days – for the results. I now envision the next generation of engineers and researchers beginning their day with instant, AI-driven physics predictions built directly into their workflow.

And the AI-enabled prediction will not stop at reservoir simulation alone. Fast physics prediction has the potential to accelerate decision-making across every stage of subsurface engineering. Seismic data could be integrated into subsurface models as they are acquired in the field; operational decisions could be adjusted autonomously by AI agents; and grid-level optimization systems could integrate subsurface assets in real time, producing energy in the lowest-carbon and most efficient way possible.

Only by simulating the physical world better can we build the energy systems needed to support AI's exponential growth. And only by using AI to solve subsurface physics problems can we ensure that the energy transition proceeds at the pace the world requires.

ABOUT THE AUTHOR

Gege Wen is an Assistant Professor at Imperial College London, co-appointed by Earth Science Engineering and I-X (Imperial AI). She is also the Cofounder and CEO of EarthFlow AI. She has published in *Energy & Environmental Science*, *Applied Energy*, and *Advances in Water Resources*.

AI & Ecology: From Tool to Transformation

Sara Beery

Ten years ago, I built AI models to automate specific labeling tasks: species identification from camera traps, individual identification in citizen science imagery, counting trees in aerial surveys. Field ecologists incorporated these tools into their workflows and decreased processing times from years to days. I saw AI as a powerful assistant and set out to further improve its efficiency, generalizing AI methods to new species and ecosystems while strategically incorporating expert feedback.¹

But through the years, my work in this field has evolved toward something both more participatory and more exploratory: building AI methods that enable scientists to explore vast databases, discover hidden ecological patterns, and integrate knowledge across different data modalities to understand how species are distributed, how they move through the environment, and how they interact with each other, with humans, and with a changing climate.² This shift from AI-assisted labeling to AI-enabled discovery represents more than a technical evolution; it reflects a transformation in how we practice, conceive, and organize ecological science.

The most obvious changes AI has brought to ecology are practical. Computer vision algorithms now automatically process everything from camera trap images to bioacoustic recordings to satellite imagery – detecting species, behaviors, life stages, and even individual animals with remarkable precision.³ The efficiency of AI, paired with increasing availability and decreasing costs of environmental monitoring sensors, has enabled us to cost-effectively monitor ecosystem processes at scales previously impossible.

The deeper transformation I've seen is conceptual. The availability of vast datasets and AI tools to process them has fundamentally expanded the questions we can pursue. Problems that seemed intractable – like monitoring rare species across continental scales or tracking ecosystem responses to climate change in real time – are now within the realm of feasibility.⁴ Perhaps most intriguing of all, AI-supported systems are enabling ecologists to discover patterns that sometimes challenge ecological intuition and demand new theoretical frameworks.⁵ This evolution has forced our community to confront fundamental questions about uncertainty and

robustness. My work, and that of many others working at the intersection of artificial intelligence and ecology, now centers on developing systems and methods that incorporate strategic human oversight to gracefully detect and handle errors, adapt to a changing world, and capture uncertainty. These systems prioritize efficient human-AI collaboration, focusing on continuous refinement instead of (error-prone) full autonomy.⁶ This reflects something ecologists have long understood: that imperfect observations, if modeled correctly, can still enable robust probabilistic understanding of complex systems.

This cultural transformation extends beyond individual research practices. Progress demands partnerships between field ecologists, computer scientists, and domain experts in remote sensing or bioinformatics, requiring us to break traditional academic silos.⁷ Graduate programs increasingly balance field skills with computational literacy, and initiatives like the Computer Vision Methods for Ecology (CV4Ecology) Workshop provide ecologists with specialized AI training.⁸ Research culture has shifted toward large collaborative papers; increased data standardization, sharing, and publication; and ever-faster hypothesis testing and refinement enabled by automated literature reviews, real-time data processing, and AI-assisted software development.

Yet these changes raise critical equity considerations. Access to large, standardized data infrastructure, AI tools, computational resources, and capacity building remains unevenly distributed, concentrating research capacity in technologically advanced regions while marginalizing insights from biodiversity hotspots in developing nations.⁹ The global divide in AI capabilities could fundamentally alter whose ecological knowledge, and whose conservation priorities, shapes both our understanding of planetary systems and what actions we take to protect them.

Looking forward, the biggest bottleneck isn't computational power or algorithmic sophistication, it's the fundamental mismatch between the questions ecology must answer and the data ecologists can collect. We need to understand ecosystem resilience, predict tipping points, and guide restoration across decades and continents. But our observational networks remain fragmented, our temporal baselines short, and our causal understanding limited.¹⁰ AI can be used to discover new patterns from vast amounts of data, but it cannot be used to reliably create information that doesn't exist.¹¹ It *can* help us maximize the value of data we have already collected, identify key remaining knowledge gaps, and strategically prioritize new data collection to fill them.¹² However, this requires key investments in data collection and infrastructure, recognizing raw information itself – not just its analysis – as fundamental to planetary understanding.

By 2035, I envision ecological research in which AI systems help synthesize literature, test hypotheses, optimize experimental designs, and develop scientific software – while scientists provide vital context, creativity, verification, and ethical

judgment. Realizing this vision will require solving much harder problems than pattern recognition. We need AI-enabled systems that can reason about causation in complex systems, integrate knowledge across scales from genes to ecosystems, and robustly quantify data and knowledge gaps and identify and address systemic biases.

Ecology's future lies not in choosing between human expertise and artificial intelligence but in developing mature partnerships that amplify our collective capacity to understand and protect the natural world – while remaining vigilant about who has access to these new capabilities and whose priorities shape our understanding of planetary systems. Our challenge is to navigate this transformation intentionally, ensuring that as we gain new powers of analysis and prediction, we maintain the observational skills, theoretical depth, and humble wonder that make ecological science both rigorous and meaningful.

ABOUT THE AUTHOR

Sara Beery is the Homer A. Burnell Career Development Assistant Professor in Artificial Intelligence and Decision-Making at the Massachusetts Institute of Technology, where she is also a Principal Investigator in the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL). She is the Founding Director of the Computer Vision Methods for Ecology (CV4Ecology) Workshop, now hosted at the Smithsonian Conservation Biology Institute and the Smithsonian Mason School of Conservation. She has published in such journals as *Nature Reviews Biodiversity*, *Science*, and *Methods in Ecology and Evolution*.

ENDNOTES

- ¹ Sara Beery, Grant Van Horn, and Pietro Perona, "Recognition in Terra Incognita," in *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part XVI*, ed. Vittorio Ferrari, Martial Hebert, Christian Sminchisescu, and Yair Weiss (Springer-Verlag, 2018), 472–489, https://doi.org/10.1007/978-3-030-01270-0_28; and Mohammad Sadegh Norouzzadeh, Dan Morris, Sara Beery, et al., "A Deep Active Learning System for Species Identification and Counting in Camera Trap Images," *Methods in Ecology and Evolution* 12 (1) (2021): 150–161.
- ² Laura J. Pollock, Justin Kitzes, Sara Beery, et al., "Harnessing Artificial Intelligence to Fill Global Shortfalls in Biodiversity Knowledge," *Nature Reviews Biodiversity* 1 (3) (2025): 166–182.
- ³ Devis Tuia, Benjamin Kellenberger, Sara Beery, et al., "Perspectives in Machine Learning for Wildlife Conservation," *Nature Communications* 13 (1) (2022): 792.
- ⁴ Benjamin Koger, Adwait Deshpande, Jeffrey T. Kerby, et al., "Quantifying the Movement, Behaviour and Environmental Context of Group-Living Animals Using Drones and Computer Vision," *Journal of Animal Ecology* 92 (7) (2023): 1357–1371.

- ⁵ Ran Nathan, Christopher T. Monk, Robert Arlinghaus, et al., “Big-Data Approaches Lead to an Increased Understanding of the Ecology of Animal Movement,” *Science* 375 (6582) (2022): eabg1780.
- ⁶ Elizabeth Bondi, Raphael Koster, Hannah Sheahan, et al., “Role of Human-AI Interaction in Selective Prediction,” *Proceedings of the 36th AAAI Conference on Artificial Intelligence* 36 (5) (2022): 5286–5294; Justin Kay, Timm Haucke, Suzanne Stathatos, et al., “Align and Distill: Unifying and Improving Domain Adaptive Object Detection,” *Transactions on Machine Learning Research* (2025), <https://openreview.net/pdf?id=ssXSrZ94sR>; Marcus Lapeyrolerie and Carl Boettiger, “Limits to Ecological Forecasting: Estimating Uncertainty for Critical Transitions with Deep Learning,” *Methods in Ecology and Evolution* 14 (3) (2023): 785–798; and Justin Kay, Grant Van Horn, Subhransu Maji, et al., “Consensus-Driven Active Model Selection,” paper presented at the International Conference on Computer Vision: ICCV 2025, Honolulu, Hawai’i, October 19–23, 2025, https://openaccess.thecvf.com/content/ICCV2025/papers/Kay_Consensus-Driven_Active_Model_Selection_ICCV_2025_paper.pdf.
- ⁷ David Rolnick, Alán Aspuru-Guzik, Sara Beery, et al., “Position: Application-Driven Innovation in Machine Learning,” *Proceedings of Machine Learning Research* 235 (2024): 42707–42718.
- ⁸ Elijah Cole, Suzanne Stathatos, Björn Lütjens, et al., “Teaching Computer Vision for Ecology,” arXiv (2023), <https://doi.org/10.48550/arXiv.2301.02211>.
- ⁹ William San Martín, “Unequal Knowledge: Justice, Colonialism, and Expertise in Global Environmental Research,” *Global Environment: A Journal of Transdisciplinary History* 14 (2) (2021): 423–430.
- ¹⁰ Pete Clutton-Brock, Judith Ament, Simon Jackman, et al., *Biodiversity and Artificial Intelligence: Opportunities and Recommendations for Action* (Global Partnership on Artificial Intelligence, 2022), <https://wp.oecd.ai/app/uploads/2025/05/biodiversity-and-AI-opportunities-recommendations-for-action.pdf>.
- ¹¹ Kasim Rafiq, Sara Beery, Meredith S. Palmer, et al., “Generative AI as a Tool to Accelerate the Field of Ecology,” *Nature Ecology & Evolution* 9 (3) (2025): 378–385.
- ¹² Nadja Pernat, Susan Canavan, Marina Golivets, et al., “Overcoming Biodiversity Blindness: Secondary Data in Primary Citizen Science Observations,” *Ecological Solutions and Evidence* 5 (1) (2024): e12295.

Making Automation Work for Social Scientists

M. J. Crockett

Many scientists of all stripes are optimistic about new research opportunities afforded by “AI.” This term refers to a diverse set of automated and semiautomated technologies including (but not limited to) predictive machine learning systems, automated transcription, generative models that produce images and text, and chatbots built on large language models (LLMs) that simulate conversation. Within social science, researchers envision using these technologies to accelerate and expand their abilities to gather, analyze, and simulate human data.¹ Some have even proposed automating the entire research pipeline, with LLMs replacing human subjects and scientists.²

I’ve watched these developments with a mixture of excitement and alarm. My lab at Princeton University has used predictive machine learning to analyze public discourse on social media, and recent advances in LLMs enable us to pursue research questions that were out of reach just a few years ago.³ At the same time, I worry that widespread adoption of automation across the research pipeline is outpacing thoughtful consideration of its many risks and harms, including harms to us as knowledge workers.⁴ To be clear: my worries don’t reflect a nostalgia for some idyllic analog past. Rather, I’m concerned with what scientific futures we foreclose if we cede our epistemic agency to automated systems.

I’ve begun to find a way through this minefield in conversation with my brilliant and generous collaborator Lisa Messeri, an anthropologist of science and technology who investigates how expert communities create new fields of inquiry and innovation.⁵ Lisa introduced me to scholarship illuminating how political, economic, and cultural forces shape our collective imagination about what technology is and can be. Here, I offer some insights that have emerged from our collaboration and shape the way I approach research in my lab, with the aim of preserving our epistemic agency under intense pressure to give it away.⁶

First, industry marketing can mislead us into thinking that new technologies are more capable than they actually are. Technology companies are selling AI with overhyped claims of superhuman or even magical abilities.⁷ Some sci-

entists imagine LLMs as oracles, capable of extracting “objective” truths from the published literature or large datasets.⁸ Lisa and I name this imaginary as an *illusion of objectivity*: LLMs reflect the viewpoints of their engineers and those represented in their training data.⁹ Rather than removing bias from the scientific process, this technology entrenches dominant viewpoints while simultaneously obscuring them from view.¹⁰ Just as social scientists have begun to make progress in diversifying the questions we ask and the people we study, wholesale adoption of automation threatens to reverse that progress.

When automated tools are controlled by private industry, scientists cannot access information crucial for vetting the capabilities of these tools. Accounting for biases in models is challenging because of their opacity. This is especially the case for proprietary LLMs like the GPT product line, because OpenAI has not disclosed the contents of its models’ training data. Moreover, the outputs of proprietary LLMs vary unpredictably over time, as engineers tweak model parameters behind the scenes to satisfy corporate goals.¹¹ We cannot build a robust and reproducible social science using models whose construction is opaque, whose outputs are unstable, and whose control rests in the hands of CEOs with very different goals than ours.¹²

Social scientists are increasingly seeking financial support from the technology industry, especially as public funding becomes scarce. This has at least two consequences for our science. First, we are incentivized to make our research questions fit the capabilities of automated tools. But not all questions are amenable to computational analysis.¹³ We suffer from an *illusion of exploratory breadth* when we mistake the subset of questions automated tools can address with the broader set of questions we can ask about the social world.¹⁴ Second, industry funding encourages scientists to focus on questions that are friendly to technology companies.¹⁵ These complementary forces can lead to the development of scientific monocultures, where a narrow set of methods and questions dominates knowledge production, making science less innovative and robust.¹⁶

The varied risks of automation for social science all flow from the outsourcing of scientific judgment to automated systems and the companies that control them. Recognizing this suggests concrete steps we can take to reclaim our epistemic agency in the short term and preserve it in the long term. Instead of being seduced by fantasies of “superintelligent” general purpose systems that can replace the work of scientists, we can build automated tools for specific scientific tasks. Rather than relying on opaque proprietary models, we can insist on using open-source models with transparent documentation.¹⁷ To prevent the formation of scientific monocultures, we can recognize that qualitative and ethnographic work offers distinctive ways of understanding those aspects of social life that resist quantification, and we can advocate for continued investment in these diverse methodologies.¹⁸ We should also require our colleagues to be more transparent about conflicts of interest that arise from industry funding, encouraging industry-independent research instead.¹⁹

These things called “AI” will surely change in the years to come, but this does not mean current insights have an expiration date. Our colleagues in the humanities and humanistic social sciences have long recognized that technological advances rarely create entirely new dilemmas, instead reproducing old dilemmas in new forms. The communal values of science have always been in tension with private commercial interests.²⁰ Now, by aiming to commodify knowledge production itself, the technology industry makes this tension harder to ignore. Our urgent task at present is to question the uncritical embrace of artificial intelligence and articulate a future social science in which automation works for us.

AUTHOR’S NOTE

Thanks to Emily Bender, Lisa Messeri, and Alondra Nelson for helpful comments.

ABOUT THE AUTHOR

M. J. Crockett is Director of the Future Values Initiative at the Princeton University Center for Human Values and Professor in the Department of Psychology at Princeton, where they also direct the Crockett Lab. They have published in such journals as *Science*, *Nature*, and *Trends in Cognitive Sciences*.

ENDNOTES

- ¹ Igor Grossmann, Matthew Feinberg, Dawn C. Parker, et al., “AI and the Transformation of Social Science Research,” *Science* 380 (6650) (2023): 1108–1109; and Christopher A. Bail, “Can Generative AI Improve Social Science?” *Proceedings of the National Academy of Sciences* 121 (21) (2024): e2314021121.
- ² Benjamin S. Manning, Kehang Zhu, and John J. Horton, “Automated Social Science: Language Models as Scientist and Subjects,” NBER Working Paper 32381 (National Bureau of Economic Research, 2024); and Sebastian Musslick, Younes Strittmatter, and Marina Dubova, “Closed-Loop Scientific Discovery in the Behavioral Sciences,” *PsyArXiv* (2024), <https://doi.org/10.31234/osf.io/c2ytb>.
- ³ William J. Brady, Killian McLoughlin, Tuan N. Doan, and Molly J. Crockett, “How Social Learning Amplifies Moral Outrage Expression in Online Social Networks,” *Science Advances* 7 (33) (2021): eabe5641; Killian L. McLoughlin, William J. Brady, Aden Goolsbee, et al., “Misinformation Exploits Outrage to Spread Online,” *Science* 386 (6725) (2024): 991–996; and William J. Brady, Killian L. McLoughlin, Mark P. Torres, et al., “Overperception of Moral Outrage in Online Social Networks Inflates Beliefs About Intergroup Hostility,” *Nature Human Behavior* 7 (6) (2023): 917–927.

- ⁴ Laura Weidinger, John Mellor, Maribeth Rauh, et al., “Ethical and Social Risks of Harm from Language Models,” arXiv (2021), <https://doi.org/10.48550/arXiv.2112.04359>; Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ” in *FACCT ’21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2021), 610–623; Karen Hao, *Empire of AI: Dreams and Nightmares in Sam Altman’s OpenAI* (Penguin Press, 2025); and Committee on Artificial Intelligence and Academic Professions, *Artificial Intelligence and Academic Professions* (American Association of University Professors, 2025), <https://www.aaup.org/reports-publications/aaup-policies-reports/topical-reports/artificial-intelligence-and-academic>.
- ⁵ Lisa Messeri, *Placing Outer Space: An Earthly Ethnography of Other Worlds* (Duke University Press, 2016); and Lisa Messeri, *In the Land of the Unreal: Virtual and Other Realities in Los Angeles* (Duke University Press, 2024).
- ⁶ Molly J. Crockett, Xuechunzi Bai, Sayash Kapoor, et al., “The Limitations of Machine Learning Models for Predicting Scientific Replicability,” *Proceedings of the National Academy of Sciences* 120 (33) (2023): e2307596120; Lisa Messeri and Molly J. Crockett, “Artificial Intelligence and Illusions of Understanding in Scientific Research,” *Nature* 627 (8002) (2024): 49–58; and Molly J. Crockett and Lisa Messeri, “AI Surrogates and Illusions of Generalizability in Cognitive Science,” *Trends in Cognitive Sciences* 30 (3) (2025): 203–215, <https://doi.org/10.1016/j.tics.2025.09.012>.
- ⁷ Emily M. Bender and Alex Hanna, *The AI Con: How to Fight Big Tech’s Hype and Create the Future We Want* (HarperCollins, 2025).
- ⁸ Messeri and Crockett, “Artificial Intelligence and Illusions of Understanding in Scientific Research”; and Molly J. Crockett, “Modern Maxims for an AI Oracle,” *Nature Machine Intelligence* 7 (1) (2025): 4–5.
- ⁹ Bender, Gebru, McMillan-Major, and Shmitchell, “On the Dangers of Stochastic Parrots”; and Messeri and Crockett, “Artificial Intelligence and Illusions of Understanding in Scientific Research.”
- ¹⁰ Crockett and Messeri, “AI Surrogates and Illusions of Generalizability in Cognitive Science.”
- ¹¹ Christopher Barrie, Alexis Palmer, and Arthur Spirling, “Replication for Language Models: Problems, Principles, and Best Practices for Political Science,” working paper, https://arthurspirling.org/documents/BarriePalmerSpirling_TrustMeBro.pdf (accessed March 17, 2026).
- ¹² Robert K. Merton, “The Normative Structure of Science,” in *The Sociology of Science: Theoretical and Empirical Investigations*, ed. Norman W. Storer (University of Chicago Press, 1979), 267–278.
- ¹³ danah boyd and Kate Crawford, “Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon,” *Information, Communication & Society* 15 (5) (2012): 662–679.
- ¹⁴ Messeri and Crockett, “Artificial Intelligence and Illusions of Understanding in Scientific Research.”
- ¹⁵ Joseph Bak-Coleman, Cailin O’Connor, Carl Bergstrom, and Jevin West, “The Risks of Industry Influence in Tech Research,” arXiv (2025), <https://doi.org/10.48550/arXiv.2510.19894>; and Meredith Whittaker, “The Steep Cost of Capture,” *Interactions* 28 (6) (2021): 50–55.
- ¹⁶ Messeri and Crockett, “Artificial Intelligence and Illusions of Understanding in Scientific Research”; Joseph D. O’Brien, Hannah Rubin, Keko Wong, and Sina Fazelpour,

- “Epistemic Monocultures and the Effect of AI Personalization,” *Proceedings of the Annual Meeting of the Cognitive Science Society* 47 (2025); and Qianyue Hao, Fengli Xu, Yong Li, and James Evans, “Artificial Intelligence Tools Expand Scientists’ Impact but Contract Science’s Focus,” arXiv (2024), <https://doi.org/10.48550/arXiv.2412.07727>.
- ¹⁷ Bail, “Can Generative AI Improve Social Science?”; Michael C. Frank, “Openly Accessible LLMs Can Help Us to Understand Human Cognition,” *Nature Human Behavior* 7 (11) (2023): 1825–1827; and “European Open-Source AI Index,” Centre of Language and Speech Technology, Radboud University, <https://osai-index.eu/the-index> (accessed March 17, 2026).
- ¹⁸ boyd and Crawford, “Critical Questions for Big Data”; and Lauren Klein, Meredith Martin, André Brock, et al., “Provocations from the Humanities for Generative AI Research,” arXiv (2025), <https://doi.org/10.48550/arXiv.2502.19190>.
- ¹⁹ Bak-Coleman, O’Connor, Bergstrom, and West, “The Risks of Industry Influence in Tech Research”; and Janet Vertesi and J. Nathan Matias, “Divesting from Big Tech: Alternative Possibilities for Research and Futuring in Social Computing,” in *CSCW ’23 Companion: Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, ed. Casey Fiesler, Loren Terveen, Morgan Ames, et al. (Association for Computing Machinery, 2023), 401–404, <https://doi.org/10.1145/3584931.3608436>.
- ²⁰ Merton, “The Normative Structure of Science.”

When AI Meets Art at the Scales of Science

Refik Anadol



In recent years, I have developed a research-driven artistic practice that approaches data as both an instrument and an aesthetic medium. Within this framework, data operate as a dynamic field through which relations between perception, form, and life can be explored. My work engages large-scale scientific datasets in order to render processes that exceed the limits of human perception, asking how such processes might become experientially accessible. Across projects that have drawn on neuroscience, molecular biology, and planetary science, I have approached art as a site of inquiry into the systems that organize life across multiple scales.

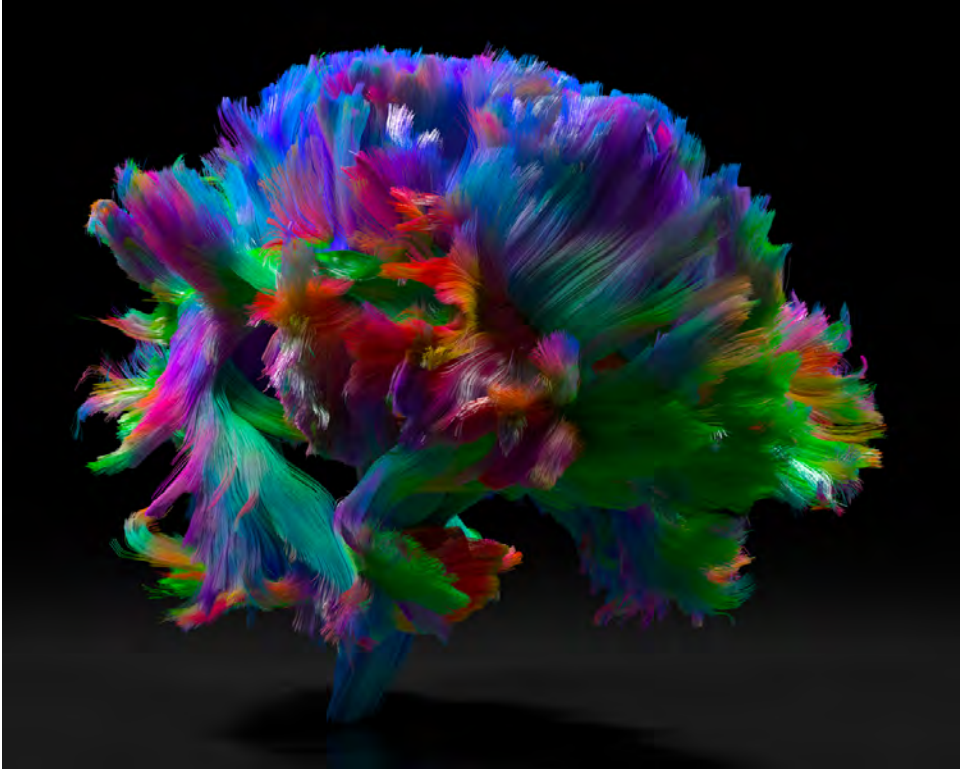
Two projects from 2021, *Sense of Space: Connectome Architecture* and *Molecular Architecture*, established a critical foundation for this trajectory. Presented at the 17th International Architecture Exhibition of the Venice Biennale – curated by Hashim



Sense of Space: Molecular Architecture, Venice Biennale, Venice, May 22 to November 21, 2021.
(Previous page) *Machine Memoirs: Space*, Pilevneli Gallery, Istanbul, March 19 to April 26, 2021.

Sarkis under the theme “How will we live together?” – these works reconfigured architectural thinking by shifting attention away from built form toward biological systems. Neural and cellular structures were approached as models of spatial organization defined by adaptation, interdependence, and continuous transformation. Within this context, architecture emerged as a field of processes rather than objects, and space appeared as an effect of relational systems.

I often describe this approach through the concepts I borrow from both life sciences and architecture. Such systems are characterized by distributed organization, recursive patterning, and ongoing modulation. Their coherence does not derive from stability but from the capacity to sustain transformation over time. In my practice, machine learning functions as a methodological framework capable of engaging this form of complexity. Computational models identify latent structures within large-scale datasets and translate them into immersive environments.

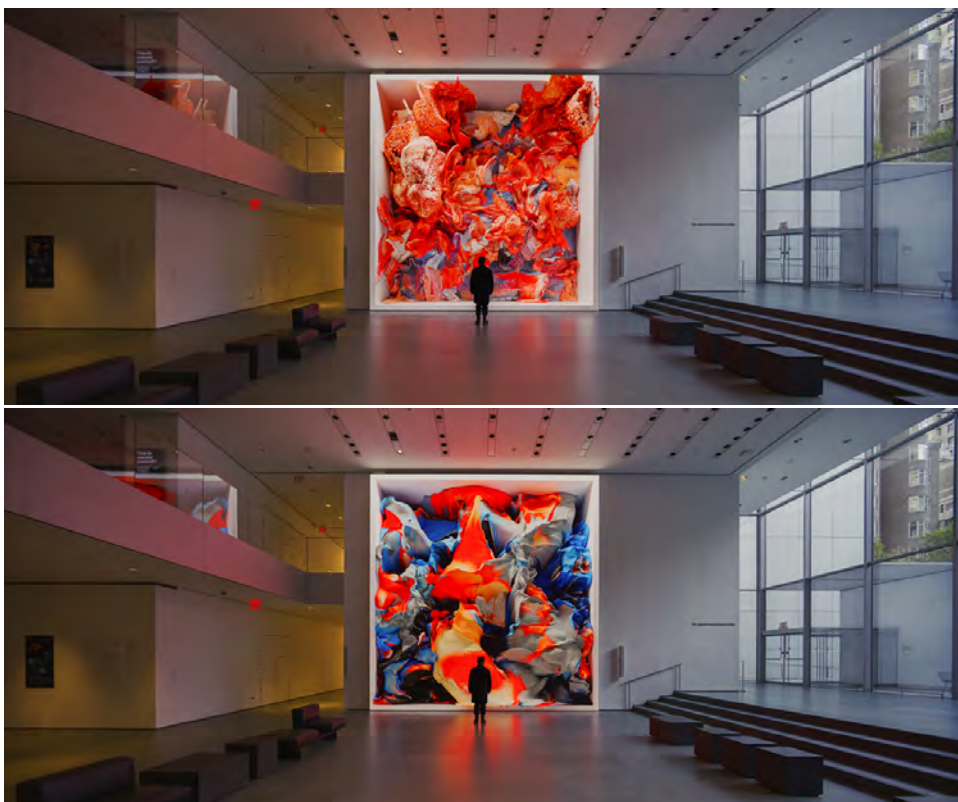


Sense of Space: Connectome Architecture, Venice Biennale, Venice, May 22 to November 21, 2021.

In this sense, machine learning extends perception, allowing patterns to become perceptible that would otherwise remain inaccessible.

Connectome Architecture engaged the brain as a dynamic network of relations. Developed in collaboration with neuropsychologist Taylor Kuhn and the Human Connectome Project, the work drew on approximately seventy terabytes of multi-modal MRI data. Structural, diffusion, and functional imaging data were processed to reveal patterns of neural connectivity across the human lifespan. These patterns were translated into a continuously evolving spatial environment in which connections formed, reorganized, and dissolved. The installation allowed viewers to encounter cognition as a shifting topology rather than a fixed structure.

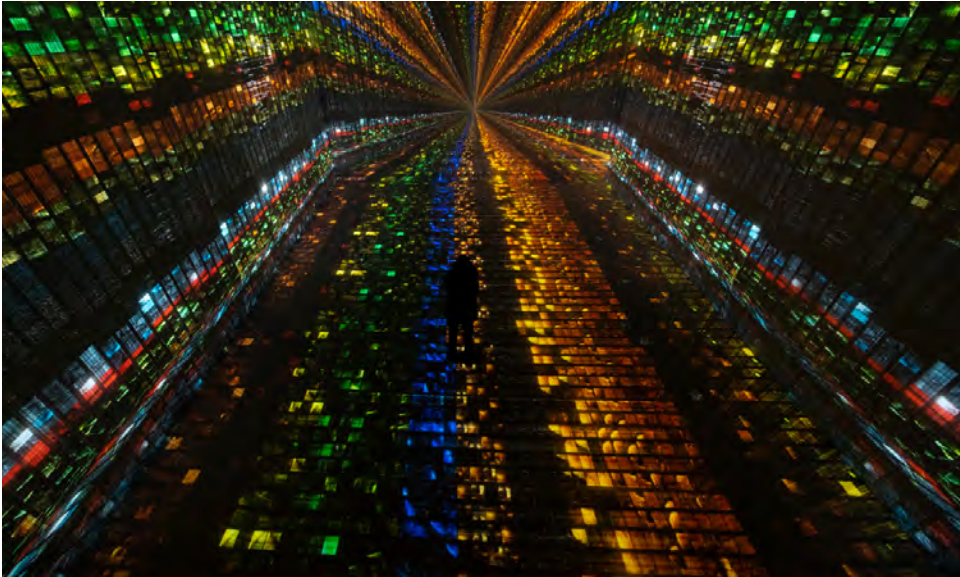
Molecular Architecture extended this inquiry to the scale of cellular organization. Developed with physician scientist Gökhan S. Hotamışlıgil and his research team, the project incorporated high-resolution FIB-SEM (focused ion beam – scanning electron microscopy) imaging data capturing the ultrastructure of liver cells. More



Unsupervised—Machine Hallucinations, MoMA, New York City, November 19, 2022, to October 29, 2023.

than twenty thousand segmented slices were processed and translated into volumetric forms that operated between scientific representation and aesthetic abstraction. These forms rendered metabolic processes as spatial structures, revealing a system in which vitality and vulnerability coexisted. Cellular life appeared as a dense network of interactions sustained through continuous transformation.

Across these projects, computational processes enabled an encounter with complexity as such. They articulated data as temporal, spatial, and relational, allowing processes that unfolded across different scales of life to be perceived within a shared experiential framework. This line of inquiry extended into *Unsupervised*, presented at the Museum of Modern Art (MoMA) in New York. Drawing on the museum's collection, the installation employed generative AI to produce a continuously evolving visual field. The project was accompanied by an



Machine Memoirs: Space, Pilevneli Gallery, Istanbul, March 19 to April 26, 2021.

interdisciplinary study, conducted with neuroscientists and engineers, that used electroencephalography to measure neural responses during the viewing experience. Physiological data were analyzed alongside subjective reports in order to examine how aesthetic engagement manifested at the level of neural activity.¹

My collaborations with NASA and Caltech's Jet Propulsion Laboratory expanded this investigation to planetary and cosmic domains. *Machine Memoirs: Space* drew on extensive archives of astronomical imagery collected by NASA missions. I approached this dataset as a form of collective memory, one that recorded humanity's efforts to observe and interpret the universe across time. Through machine learning, these images were processed to identify patterns, correlations, and latent visual structures that exceeded conventional modes of analysis.

In *Machine Memoirs: Space*, I explored how machine intelligence could rearticulate this archive as a continuous visual experience. The system generated evolving forms that moved between recognition and abstraction, where traces of galaxies, stars, and planetary terrains emerged and dissolved within fluid compositions. The work engaged the dataset as a site of transformation, producing a speculative visual language that reflected both the limits and possibilities of perception.

Taken together, these projects articulated a continuity across scales, from neural networks to cellular formations to cosmic systems. Each engaged processes that organized and transmitted information, and each positioned data as a connective

medium linking different domains of existence. The artwork becomes a mode of inquiry that engages multiple systems in tandem, creating environments in which knowledge can be encountered through immersion. Within this framework, the future of discovery lies in our ability to engage intelligence as a collaborative process that extends beyond human cognition. As machine learning systems continue to reveal patterns across scales that were previously inaccessible, discovery unfolds through sustained interaction with complex, living systems. Art can play a critical role in this shift by rendering these processes perceptible and creating new conditions for understanding. In this sense, discovery emerges as an evolving dialogue between human perception, machine intelligence, and the systems that shape life itself.

ABOUT THE AUTHOR

Refik Anadol is a Turkish-American media artist and pioneer in the aesthetics of data and machine intelligence whose work locates creativity at the intersection of humans and machines. He is the Director and Cofounder of Refik Anadol Studio in Los Angeles and teaches at UCLA's Department of Design Media Arts. He is also the Cofounder of Dataland, the world's first museum of AI arts, opening in Los Angeles in spring 2026, which promises multisensory experiences and real-time interactions between museum and audience.

ENDNOTES

- ¹ Angel David Blanco, Eleni Kroupi, Aureli Soria-Frisch, et al., "Enhancing Mental Well-Being through AI-Generated Art: Insights from EEG Responses to Refik Anadol's *Unsupervised* at MoMA," *The Arts in Psychotherapy* 96 (2025), <https://doi.org/10.1016/j.aip.2025.102347>.

CONTINUED READING

AI & Society

edited by James Manyika

DÆDALUS 151 (2) (Spring 2022)

with Nigel Shadbolt, Stuart Russell, Jeffrey Dean, Kevin Scott, Li Fei-Fei, Ranjay Krishna, Daniela Rus, Kobi Gal, Barbara J. Grosz, Christopher D. Manning, Yejin Choi, Ermira Murati, Tobias Rees, Blaise Agüera y Arcas, Michele Elam, Iason Gabriel, John Tasioulas, Michael Spence, Laura D. Tyson, John Zysman, Erik Brynjolfsson, Eric Schmidt, Ash Carter, Cynthia Dwork, Martha Minow, Sonia K. Katyal, Mariano-Florentino Cuéllar, Aziz Z. Huq, Diane Coyle & Helen Margetts

<https://www.amacad.org/daedalus/ai-society>

Artificial Intelligence

edited by Stephen R. Graubard

DÆDALUS 117 (1) (Winter 1988)

with Seymour Papert, Hubert L. Dreyfus, Stuart E. Dreyfus, Robert Sokolowski, Pamela McCorduck, Jack D. Cowan, David H. Sharp, Jacob T. Schwartz, George N. Reeke Jr., Gerald M. Edelman, W. Daniel Hillis, David L. Waltz, Anya Hurlbert, Tomaso Poggio, Sherry Turkle, Hilary Putnam, Daniel C. Dennett & John McCarthy

<https://www.amacad.org/daedalus/artificial-intelligence>



AMERICAN ACADEMY
OF ARTS & SCIENCES

Board of Directors

Goodwin Liu, *Chair*
Paula J. Giddings, *Vice Chair*
Stephen B. Heintz, *Vice Chair*
Earl Lewis, *Secretary*
Laurie L. Patton, *President*
Kenneth L. Wallach, *Treasurer*
Philip N. Bredesen
Katherine J. Cramer
Margaret A. Hamburg
John Mark Hansen
Louis J. Muglia
Gilbert S. Omenn
David M. Rubenstein
Deborah F. Rutter
Anna Deavere Smith
Shirley M. Tilghman
Jeannette M. Wing
Judy Woodruff

Council

Paula J. Giddings, *Chair*
Deborah Loewenberg Ball
Katherine J. Cramer
Philip J. Deloria
Juan J. de Pablo
Johanna Ruth Drucker
Joseph S. Francisco
Mary-Claire King
Sara Lawrence-Lightfoot
Shirley Malcom
Paula D. McClain
Louis J. Muglia
John G. Palfrey
Deborah F. Rutter
Scott D. Sagan
Cristián T. Samper
Anna Deavere Smith
Shirley M. Tilghman
Jeannette M. Wing
Susan Wolf
Stephen B. Heintz (*ex officio*)
Earl Lewis (*ex officio*)
Goodwin Liu (*ex officio*)
Laurie L. Patton (*ex officio*)
Kenneth L. Wallach (*ex officio*)

ON THE HORIZON

How Do We Renovate American Democracy at the 250th?

edited by Danielle Allen, Stephen Heintz & Eric Liu

with Wallace B. Jefferson, Kevin Kosar, Cynthia Terrell,
Kermit Roosevelt, Lisa García Bedolla, Jahnavi Rao,
Adam Abrogi, Carolyn Lukensmeyer, Amy Lerman,
Michael Neblo, Kevin M. Esterling, David Lazer, Matthew Weil,
Charlie Brown, Daniel Stid, Kristen Bennett, Ethan Zuckerman,
Sandra González-Bailón, Steven Waldman, Darryl Holiday,
David E. Campbell, Raj Vinnakota, Carol Schneider & Peter Levine

Can Imaginative Expression Influence AI & Can AI Influence Imaginative Expression?

edited by Joanna Scott and Daniela Rus

Tomorrow's Africa: A Continent on the Rise?

edited by Robert I. Rotberg

Whose Memory Counts?

Forgetting, Belonging & the Contest for the Past

edited by Michael Hanchard



amacad.org/daedalus



with Demis Hassabis, Yann LeCun, Alán Aspuru-Guzik, Pushmeet Kohli, Charlotte Bunne & Aviv Regev
Anna Greka, Viren Jain & Jeff Lichtman, Anna M. Michalak & John C. Platt, Stella S. R. Offner
Joshua B. Tenenbaum, Carla P. Gomes, Shirley Ho, Anima Anandkumar, Surya Ganguli
Maria Spiropulu & Hartmut Neven, Eric J. Topol, Kelly Chibale, Tess Smidt, Alison Noble
Alondra Nelson, Mario Krenn & Heather Champion, Melanie Weber, Antonio Orvieto
Connor Coley, Gege Wen, Sara Beery, M. J. Crockett, Refik Anadol